

Technical Report 417

**LEVEL**

12

**STATISTICAL MODELS FOR  
CRITERION-REFERENCED TESTING AND  
DECISIONMAKING**

Kenneth I. Epstein

**SIMULATION SYSTEMS TECHNICAL AREA**

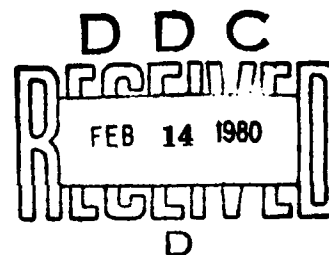


**U. S. Army**

**Research Institute for the Behavioral and Social Sciences**

**October 1979**

Approved for public release; distribution unlimited.



DDC FILE COPY

ADA080651

# U. S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the  
Deputy Chief of Staff for Personnel

**JOSEPH ZEIDNER**  
Technical Director

**WILLIAM L. HAUSER**  
Colonel, U S Army  
Commander

---

## NOTICES

**DISTRIBUTION:** Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

**FINAL DISPOSITION:** This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

**NOTE:** The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report 417	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) STATISTICAL MODELS FOR CRITERION-REFERENCED TESTING AND DECISION MAKING	5. TYPE OF REPORT & PERIOD COVERED Final Technical Report	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Kenneth I. Epstein	8. CONTRACT OR GRANT NUMBER(s)	
9. PERFORMING ORGANIZATION NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue, Alexandria, VA 22333	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q762722A764	
11. CONTROLLING OFFICE NAME AND ADDRESS Deputy Chief of Staff for Personnel Washington, DC 20310	12. REPORT DATE October 1979	13. NUMBER OF PAGES 216
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report)  Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)  ---		
18. SUPPLEMENTARY NOTES  This publication is based upon a dissertation submitted by Kenneth I. Epstein in partial fulfillment of the requirement for the Ph.D.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)  Criterion-referenced testing Statistical models Performance evaluation		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)  The purposes of this study were to investigate the characteristics of a well-constructed criterion-referenced performance test and to compare several statistical models that might help interpret criterion-referenced test scores. The models were compared on the accuracy of the pass or fail decisions which they implied and the accuracy of their estimates of examinee true scores. The data base consisted of scores achieved by U.S. Army Military Police trainees on the Military Police Firearms Qualification Course, a criterion-referenced (Continued)		

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 68 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

408 010

next  
page  
JCB

## Item 20 (Continued)

performance test designed to assess .45 caliber pistol marksmanship skills. Trainees fired 240 rounds apiece to define their pass or fail classification and their true ability. Subtests of 10, 20, 40, 80, and 120 rounds were also sampled.

The three models that were considered share the binomial probability distribution for describing the expected distribution of observed scores given an examinee's true ability, and all define ability on a scale from 0 to 1.0. The first model, the proportion correct model, uses the proportion of responses that are correct as its estimate of true ability. Pass or fail criteria are set by considering the probabilities that examinees of differing abilities will achieve a variety of proportion correct scores. The score that would be expected to produce the least amount of classification error is chosen as the criterion score. The second model, the binomial error model, uses the observed score distribution to compute the regression of true score on observed score. Pass or fail decisions and true score estimates are based on the results of applying the regression equation. The third model, the beta-binomial Bayesian model, uses prior beliefs of expert judges to establish a prior ability distribution. Observed data are combined with the prior distribution to produce a posterior ability distribution for each observed score. Pass or fail decisions and true score estimates are based on the posterior distributions.

Criterion-referenced tests can be evaluated by a variety of logical and empirical analyses. The analyses include descriptions of the skill domain, the rationale for choosing test items, the purposes of the test, the level of skill chosen to represent adequate skill mastery, and the expected results of administering the test to specified groups of examinees. Descriptive and inferential statistical techniques can empirically confirm or question the logical analysis of a criterion-referenced test.

The comparison of the statistical models indicated relatively few differences between the models and no evidence that one was better or worse than others. The comparison data did, however, clearly demonstrate the importance of a close match between test items and the domain to which results are to be generalized. When test items did not match the skill domain, the risk of incorrect classification decisions was high, the magnitude of the decision errors was not accurately predicted by statistical considerations, and the true abilities of examinees were poorly estimated by all of the models. When the items more closely approximated the domain, the amount of classification error decreased and became more predictable, and true abilities were more accurately estimated.



# STATISTICAL MODELS FOR CRITERION-REFERENCED TESTING AND DECISIONMAKING

Kenneth I. Epstein

Angelo Mirabella, Team Chief

Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DDC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/_____	
Availability Codes	
Dist.	Avail and/or special
A	

Submitted by:  
Frank J. Harris, Chief  
SIMULATION SYSTEMS TECHNICAL AREA

Approved by:

Milton S. Katz, Acting Director  
ORGANIZATIONS AND SYSTEMS  
RESEARCH LABORATORY

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES  
5001 Eisenhower Avenue, Alexandria, Virginia 22333

Office, Deputy Chief of Staff for Personnel  
Department of the Army

October 1979

Army Project Number  
2Q762722A764

Education & Training

Approved for public release; distribution unlimited.

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

---

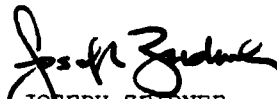
## FOREWORD

---

The research presented in this report was conducted under Project METTEST (Methodological Issues in Criterion-Referenced Testing), under the auspices of the Engagement Simulation Technical Area of the Army Research Institute for the Behavioral and Social Sciences (ARI), and under Army Project 2Q762722A764. The goal of Project METTEST has been to develop quantitative methods for evaluating unit proficiency. The means for achieving this goal include basic research in test construction, measurement and decisionmaking models, and computer-programmable models for large-scale data analysis.

This report uses data from an earlier investigation of the Military Police Firearms Qualification Course, described in ARI Technical Paper 322, to compare the usefulness of several standard statistical models in evaluating and interpreting criterion-referenced test scores.

Related programs within the technical area have included evaluation of small combat units under simulated battlefield conditions (REALTRAIN, ARTEP), qualification of tank gunnery crews and revision of table VIII (IDOC), and combat effectiveness evaluation by group decision making and board-game simulation (COTEAM, or Combat Operations Training Effectiveness Analysis).

  
JOSEPH ZEIDNER  
Technical Director

## STATISTICAL MODELS FOR CRITERION-REFERENCED TESTING AND DECISIONMAKING

### BRIEF

---

#### Requirement:

To describe the operating characteristics of a well-constructed criterion-referenced performance test and to compare the potential usefulness of several statistical models in helping to interpret criterion-referenced test scores. The models were compared on the basis of the accuracy of pass/fail decisions which they implied and accuracy of their estimates of examinees' true scores.

#### Procedure:

A criterion-referenced performance test of pistol marksmanship, the Military Police Firearms Qualifications Course (MPFQC), was evaluated on logical and empirical grounds. The evaluation included description of the skill domain, rationale for choosing test items, purposes of the test, level of skill chosen to represent adequate skill mastery, and expected results of administering the test to specified groups of examinees. Test scores which military police trainees obtained on the MPFQC were then used as a data base for comparing three statistical models: the proportion correct model, the binomial error model, and the beta-binomial Bayesian model.

#### Findings:

Descriptive statistics and inferential techniques such as means, variances, and analysis of variance can empirically confirm or indicate error in the interpretation of the logical analysis of a criterion-referenced test. Logical analysis indicated that the MPFQC fulfilled the requirements for a well-designed criterion-referenced performance test. Empirical analysis indicated, not the assumed unitary skill domain, but a two-dimensional domain and suggested that test scores could be interpreted either in terms of the overall domain or independently for each of two subdomains.

Comparison of the statistical models indicated relatively few practical differences among them and no evidence that one was better or worse than the others. The comparison data did, however, clearly demonstrate the importance of a close match between test items and the skill domain being tested. When test items did not match the domain, the risk of incorrect classification decisions was high, the size of decision errors was not accurately predicted statistically, and all the models did poorly in estimating examinees' true abilities. When the items more closely approximated the domain, classification error decreased and became more predictable, and true abilities were more accurately estimated.

#### Utilization of findings:

Decision errors will probably always be a problem when criterion-referenced tests are administered. The most important action that can be taken to keep decision error to a reasonable level is to insure that the test items adequately represent the skill domain they are intended to measure. If the match between test items and domain is good, then statistical models can be used along with subjective estimates of the proportion of masters to nonmasters in the examinee group to estimate the types and amounts of misclassification error and its impact on decisionmaking.

STATISTICAL MODELS FOR CRITERION-REFERENCED TESTING AND DECISION-  
MAKING

TABLE OF CONTENTS

ABSTRACT . . . . .	ii
ACKNOWLEDGEMENTS . . . . .	iii
LIST OF TABLES . . . . .	ix
LIST OF FIGURES . . . . .	x
. . . . .	
Chapter	
1. INTRODUCTION . . . . .	1
Statement of the Problem	
2. REVIEW OF THE LITERATURE . . . . .	3
Criterion-referenced Testing	
True Score and Human Capabilities	
Measurement Models	
3. METHODS. . . . .	42
The Data Base	
Test Characteristics	
Measurement Models	
Comparing the Models	
4. RESULTS. . . . .	76
Characteristics of the MPFQC Performance Data	
Comparison of the Scoring Models: 240 Round Criterion	
Comparison of the Scoring Models: 120 Round Hard and	
Easy Criteria	
5. DISCUSSION . . . . .	145
Analysis of the MPFQC	
Comparison of the Models	
6. SUMMARY AND CONCLUSIONS . . . . .	166
APPENDIX . . . . .	169
REFERENCES . . . . .	210

# LIST OF TABLES

Table 1.	True Mastery State and Measurement Error for the Emrick Model . . . . .	20
Table 2.	MPFQC Shot Groups, Tables, and Sampled Subtests; Means and Reliabilities . . . . .	46
Table 3.	Proportion Correct Model Probabilities of False Positive and False Negative Misclassification Errors for a Variety of Test Lengths, Criterion Scores, and True Abilities . . . . .	53
Table 4.	Binomial Error Model $\chi^2$ Probabilities that Subtest Scores Represent Samples from a Negative Hypergeometric Distribution, and Criterion Observed and Estimated True Scores . . . . .	62
Table 5.	Beta-Binomial Bayesian Model Probabilities that Ability $\geq .70$ as a Function of Observed Score and Prior Distribution . . . . .	66
Table 6.	Expected Examinee Performance on the Military Police Firearms Qualification Course and Implied Prior Beta Distributions . . . . .	71
Table 7.	Analysis of Variance Summary Table and Proportion of Total Variance Accounted for by Main Effects and Interactions . . . . .	91
Table A.	Recommended Criterion Scores and Observed, Expected, and Observed versus Expected Misclassification Rates: 240 Round Criterion . . . . .	170
Table B.	Average Per Test Sum of Squared and Absolute Discrepancies Between Estimated True Scores and Criterion True Scores: 240 Round Criterion. . . . .	188
Table C.	Recommended Criterion Scores and Observed, Expected, and Observed versus Expected Misclassification Rates: 120 Round Hard and Easy Criteria. . . . .	194
Table D.	Average Per Test Sum of Squared and Absolute Discrepancies Between Estimated True Scores and Criterion True Scores: 120 Round Hard and Easy Criteria . . . . .	206

## LIST OF FIGURES

Figure 1.	The Military Police Firearms Qualification Course. . . .	44
Figure 2.	True Classification versus Subtest Classification Contingency Matrix . . . . .	73
Figure 3.	Distribution of Scores on 240 Round Criterion Test . . .	77
Figure 4.	Distribution of Scores on 80 Round Subtests. . . . .	79
Figure 5.	Distribution of Scores on 120 Round Hard Criterion Test . . . . .	80
Figure 6.	Distribution of Scores on 120 Round Easy Criterion Test . . . . .	81
Figure 7.	Test Characteristic Curves . . . . .	84
Figure 8.	Observed Misclassification Rates: 240 Round Criterion . . . . .	99
Figure 9.	Scatterplot of Test Difficulty and Best Criterion Score When Best Criterion Score is Defined as the Score Producing the Lowest Total Misclassification Compared to the 240 Round Criterion . . . . .	104
Figure 10.	Scatterplot of Test Difficulty and Best Criterion Score When Best Criterion Score is Defined as the Score Producing the False Positive to False Negative Misclassification Rate Ratio Closest to 1.0 Based on 240 Round Criterion . . . . .	105
Figure 11.	Expected Misclassification Rates: 240 Round Criterion . . . . .	106
Figure 12.	Absolute Values of Differences Between Observed and Expected Misclassification Rates: 240 Round Criterion . . . . .	113
Figure 13.	Average Sum of Squared Discrepancies: Subtest Estimated True Scores and 240 Round Criterion . . . . .	122



Figure 14.	Average  Sum of Absolute Discrepancies : Subtest Estimated True Scores and 240 Round Criterion. . . . .	123
Figure 15.	Observed Misclassification Rates: Hard Subtests and 120 Round Hard Criterion . . . . .	131
Figure 16.	Observed Misclassification Rates: Easy Subtests and 120 Round Easy Criterion . . . . .	132
Figure 17.	Expected Misclassification Rates: Hard Subtests and 120 Round Hard Criterion . . . . .	135
Figure 18.	Expected Misclassification Rates: Easy Subtests and 120 Round Easy Criterion . . . . .	136
Figure 19.	Absolute Values of Differences Between Observed and Expected Misclassification Rates: Hard Sub- tests and 120 Round Hard Criterion . . . . .	139
Figure 20.	Absolute Values of Differences Between Observed and Expected Misclassification Rates: Easy Sub- tests and 120 Round Easy Criterion . . . . .	140
Figure 21.	Average Sum of Squared Discrepancies and Average  Sum of Absolute Discrepancies : Subtest Estimated True Scores and 120 Round Hard Criterion . .	142
Figure 22.	Average Sum of Squared Discrepancies and Average  Sum of Absolute Discrepancies : Subtest Estimated True Scores and 120 Round Easy Criterion . .	142

## STATISTICAL MODELS FOR CRITERION-REFERENCED TESTING AND DECISIONMAKING

---

### 1. INTRODUCTION

The growing acceptance of instructional systems technology and the widespread use of objectives in education make it critical that measurement techniques responsive to the needs of objectives based instruction be investigated. The heavy investment in time and money required for the development of instructional systems does not allow for casual testing programs. Decisions concerning students' abilities, needs, and advancement opportunities must be based on valid and reliable data. One attempt to meet the need for a strong measurement component in instructional systems technology lies in the field of criterion-referenced measurement.

Criterion-referenced measurement provides data which are interpreted in terms of examinees' abilities to achieve an objective or to do a task. Decisions are based on how well they perform. Often the decision making process will collapse to a simple dichotomy; students pass or fail, they are masters or nonmasters, they are promoted to the next unit of instruction or recycled for remedial work.

Unfortunately, even very good criterion-referenced tests are not error free. Items may not adequately reflect the objectives or tasks for which criterion-referenced tests are designed, leading to problems of test validity. Whether or not a test is valid, observed performance incorporates some degree of error inherent in the measurement process itself. In order to help interpret the fallible observed scores,

measurement models are developed to estimate the value of the error free true score that corresponds to an observed score, to support the decision making process based on the fallible observed scores directly, or both. The purpose of this research is to compare several measurement models that may be applicable to criterion-referenced testing in terms of the accuracy of their true score estimates and their implications for dichotomous decision making.

#### Statement of the Problem

Criterion-referenced tests are designed to provide data to support decisions relating to a student's ability to perform the tasks described by a well defined objective or skill domain. The items included on criterion-referenced tests are assumed to be relatively homogeneous with respect to both content and difficulty. Measures of ability obtained through criterion-referenced testing should be stable and accurate. However, the process of measurement involves error. Measurement models are designed to improve decision making by mathematically defining the measurement process and by specifying procedures which allow inferences based on observed data to be made with minimum amounts of error. The estimates of examinee error free true scores, and, in some cases, the decisions that are made concerning examinees' abilities will vary for different measurement models. The purpose of this study is to describe the operating characteristics of one criterion-referenced test, to compare several measurement models on theoretical and empirical grounds, and to suggest guidelines for choosing a model for a given testing situation.

## 2. REVIEW OF THE LITERATURE

### Criterion-referenced Testing

#### Definitions

The literature on criterion-referenced testing (CRT) is extensive and is characterized by a proliferation of definitions. For example, Donlon (1974) pointed out that by the fall of 1973, over 350 references were known by the ERIC Center on Tests, Measurement and Evaluation. He also listed ten alternative terms for score referencing, eight of which can be interpreted as special cases of criterion-referenced testing. More recently, Hambleton, Swaminathan, Algina, and Coulson (1978) noted that the number of references has increased to over 600.

The generally acknowledged first use of the term "criterion-referenced" is in a 1963 article by Robert Glaser. In that article Glaser wrote, "Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement" (p.520). The most important feature of Glaser's definition, that the "content of the behavioral repertory" is being measured by a criterion-referenced test, seems to have endured. The major controversy seems to lie in how the tester insures that a test does relate to the "behavioral repertory", and in how to interpret the "correspondence between what an individual does and the underlying continuum of achievement". Thus terms such as "content standard score" (Ebel, 1962), "universe-defined tests"

(Hively, Patterson, & Page, 1968), "Domain-referenced test" (Millman, 1973), and "objectives-based tests" (Baker, 1974) have appeared in the literature in various authors' attempts to make Glaser's basic conceptualization more concrete and usable. The most recent work concerning test specifications is described in papers by Popham (1978) and Millman (1978).

This study is primarily concerned with models for interpreting the results of criterion-referenced tests for decision making. The author of each model considered presents a unique definition of a criterion-referenced test. However, a common characteristic of all the definitions included in the models, as well as the alternative terms suggested above, does exist. Davis (1972) has specified this common characteristic, "In constructing a criterion-referenced test, the behavior categories that are to be measured must be clearly specified in a test outline. Items are then devised to test these behaviors" (p.1). In choosing the items that are to be included in a single test or subtest, Davis further suggests that they be "homogeneous in the sense that they test performance on one specific behavior or cluster of behaviors" (p.12). For purposes of this study, any test that satisfies Davis' guidelines will be considered a criterion-referenced test.

#### True Score

Interpreting an individual's performance on a CRT in terms of the underlying continuum of achievement presents further problems. Regardless of how carefully a test designer specifies the behavior to be observed and prepares test items or exercises that correspond to the

specified behavior, the observed performance is subject to uncertainty. Thus, some procedure must be available for translating the observed score into the score that would be obtained were the test free of measurement error, the true score. The manner in which a particular CRT model defines or conceptualizes true score forms one important distinguishing characteristic of the model.

Roudabush (1974) points out the importance of the definition of the true score and suggests two models describing the underlying nature of the attribute being measured by a CRT. "The first assumes an underlying all-or-none, dichotomous, 'true' score and the second assumes an underlying continuous 'true' score" (p.4). The choice of the type of true score being estimated has important implications for the interpretation of both a given observed score and the nature of the error. For example, assume that a measurement procedure is developed to assess an individual's ability to perform a particular task. If an individual performed the task 100 times, 85 times correctly and 15 times incorrectly, how could these observations be interpreted? If the continuous true score model is assumed, one might say that the observed score of 85 correct is an unbiased estimate of an individual's ability characterized by an expected proportion correct of 0.85 over all possible task administrations. Depending on distributional assumptions, one could then calculate the probability of obtaining an observed score of 85 correct in 100 trials given a true ability of 0.85.

Under the all-or-none true score model an individual can only be a "true" all correct type or a "true" none correct type. Thus, if an observed score of 85 was obtained, one might infer that the individual obtaining that score was a "true" all correct type who responded to

this particular fallible measure with a 15% error rate. Alternatively, one might infer that the individual was a "true" none correct type who responded to this particular fallible measure with an 85% luck guess rate. Under certain distributional assumptions, the probabilities of a "true" all correct type and a "true" none correct type obtaining a score of 85 correct out of 100 trials could be calculated.

Three inferences about an individual for whom 85 correct responses are observed in 100 trials are suggested. The individual could have a true ability estimated as 0.85; he or she could be a "true" all correct type who committed 15 errors; or he or she could be a "true" none correct type who made 85 lucky guesses. In this case, the distinction between the all-or-none and the continuous true score models may be trivial. Unless the measure approaches uselessness, it is highly unlikely that 85 correct responses in 100 trials would be achieved by a "true" none correct type. Further, the difference between a "true" all correct type and a true 0.85 type is marginal and unlikely to be of importance except for highly critical tasks, or when an exceptionally high level of precision is required. However, consider the case of observing 50 successes in 100 trials. In this case the choice of the model becomes critical for any interpretation to be meaningful. A "true" all correct type who happened to have made 50 errors is quite different from a "true" 0.50 type. For the "true" all correct interpretation these results describe a rather careless individual who should be allowed to continue with the next unit of instruction. However, under the continuous model these results would probably indicate an individual who has not adequately mastered the instruction and who

needs considerable remedial work.

### Decision Making

Assumptions regarding the nature of true scores along with the type of measurement procedure used have implications for the calculation of decision making error. Roudabush (1974) considered four cases:

- Case I: a dichotomous measure of a dichotomous true score;
- Case II: a pseudo continuous measure of a dichotomous true score;
- Case III: a dichotomous measure of a continuous true score; and
- Case IV: a pseudo continuous measure of a continuous true score.

For Case I, misclassification errors occur when "true" all correct types incorrectly respond to the measure, and when "true" none correct types correctly respond to the measure. The probability of misclassification can be calculated according to the following equation:

$$P(m) = P(X=1|T=0) + P(X=0|T=1),$$

where  $P(m)$  is the probability of misclassification,  $P(X=1|T=0)$  is the probability that "true" none correct types respond correctly, and  $P(X=0|T=1)$  is the probability that "true" all correct types respond incorrectly.

For Case II a complication arises. The pseudo continuous nature of the measure implies that scores may take values from 0 to  $n$ , where  $n$  is the maximum possible score. Therefore, a score between 0 and  $n$  must be defined as the minimum observed score required for an individual to "pass". Common values for the minimum score, which will be referred to as the criterion score,  $X_c$ , are the nearest integer value corresponding to  $0.80n$ ,  $0.85n$ , or  $0.90n$ . Misclassification errors under Case II can



occur when "true" all correct types obtain a score below the criterion score. The equation for calculating the probability of misclassification is

$$P(m) = P(X \geq X_c | T=0) + P(X < X_c | T=1),$$

where  $P(m)$  is the probability of misclassification,  $P(X \geq X_c | T=0)$  is the probability that "true" none correct types obtain a score at or above the criterion score, and  $P(X < X_c | T=1)$  is the probability that "true" all correct types obtain a score below the criterion score.

Cases III and IV require that a criterion true ability be defined. The criterion true ability may be thought of as the minimum true ability required for an individual to be considered capable. The criterion true ability will be denoted  $A$ . Case III applies to a dichotomous measure of a continuous true score. Misclassification errors occur when individuals of ability greater than or equal to the criterion true ability incorrectly respond to the measure and when individuals of ability below the criterion true ability respond correctly. The probability of misclassification is

$$P(m) = P(X=1 | T < A) + P(X=0 | T \geq A),$$

where  $P(m)$  is the probability of misclassification,  $P(X=1 | T < A)$  is the probability that individuals of true ability below the criterion true ability respond correctly, and  $P(X=0 | T \geq A)$  is the probability that individuals of true ability at or above the criterion true ability respond incorrectly.

Case IV is the most complex of the situations discussed. It calls for the definition of both a criterion true ability and a criterion score. Misclassification errors occur when individuals of true ability

at or above the criterion true ability obtain observed scores below the criterion score and when individuals of true ability below the criterion ability obtain scores at or above the criterion score. The probability of misclassification is

$$P(m) = P(X \geq X_c | T < A) + P(X < X_c | T \geq A),$$

where  $P(m)$  is the probability of misclassification,  $P(X \geq X_c | T < A)$  is the probability that individuals of true ability below the criterion true ability obtain scores at or above the criterion score, and  $P(X < X_c | T \geq A)$  is the probability that individuals of ability at or above the criterion true ability obtain scores below the criterion score.

The value of  $P(m)$  will vary depending on which case applies. Thus decision makers must consider their assumptions concerning the nature of whatever it is they are measuring in order for interpretations to be meaningful. In fact, the value of any decision making rule may be questionable if logical or empirical analysis of the measurement procedure and the property or attribute being measured indicates that the underlying model is inappropriate.

This discussion has not addressed the relative costs of misclassification. That is, it has been tacitly assumed that whatever losses occur as a result of incorrectly classifying a master as a nonmaster are equivalent to those resulting from the incorrect classification of a nonmaster as a master. A number of authors (e.g., Block, 1972; Hambleton and Novick, 1973; Novick and Lewis, 1974; Hambleton, Swaminathan, Algina, & Coulson, 1978) have criticized this assumption and suggested procedures to deal with unequal losses. The problem is not addressed in this study for two reasons. First, each of the models

considered could be elaborated to include the relative costs of misclassification. However, this would complicate the implementation and discussion of the models without substantially contributing to the comparison. Second, although misclassification costs may differ greatly in certain applications, particularly those involving certification or licensing, they appear to be ignored in most instructional programs implementing criterion-referenced tests (Hambleton, Swaminathan, Algina, & Coulson, 1978).

#### True Score and Human Capabilities

In order to choose a measurement model to evaluate test results, knowledge concerning the nature of the attribute being measured must be available. Gagné and Briggs (1974) present a framework for research with the potential for supplying the information necessary to choose an appropriate measurement model. Human capabilities are divided into five general categories in the Gagné and Briggs model: intellectual skills; cognitive strategies; information; attitudes; and motor skills.

#### Intellectual Skills

Intellectual skills allow an individual to deal with conceptualizations and relationships within his environment. They can be as simple as discriminating between two different geometric figures, or as complex as deriving a system of relationships to explain the workings of society. Evidence that an intellectual skill has been acquired is shown when "it is possible to say with confidence that the learned performance has a kind of 'regularity' over a variety of specific situations. In other words, the learner shows that he is able to respond with a class of relationships among classes of objects and events"

(Gagné and Briggs, 1974, p.43). This implies that performance of an intellectual skill is expected to be displayed in an "all-or-none" fashion, or that measurement is over a true dichotomous variable. Either the individual has the capability implied by the skill, in which case it can be applied repeatedly, or the skill has not been learned, in which case the individual would not be expected to be able to apply it. Inconsistent behavior may imply that solutions to specific problems have been memorized as opposed to acquisition of the necessary intellectual skill. Graham (1974) and Graham and Bergquist (1975) report studies which indicate that tests designed to measure acquisition of unitary, explicitly defined intellectual skills yield essentially bimodal distributions, demonstrating the viability of the dichotomous variable assumption.

#### Cognitive Strategies

Intellectual skills provide a means for the individual to deal with objects and relationships in his environment. By contrast, cognitive strategies refer to the individual's own internal thought processes. In other words, cognitive strategies are the skills that are used to organize and guide the internal processes involved in defining and solving novel problems. Evidence for the acquisition of cognitive strategies is shown when the individual is able to develop solutions to problem situations "in which neither the class of solution nor the specific manner of solution are specified for the learner. The learner needs to have available a variety of cognitive strategies of problem solution from which he can make a selection" (Gagné and Briggs, 1974, p.49).

The measurement procedures required for determining the acquisition of cognitive strategies present problems quite different from those for intellectual skills. Precise operational definitions of cognitive strategies have yet to be developed. Further, a taxonomy of cognitive strategies, which would allow for determination of the specific skill or skills used in solving unique problems, is not yet available. At this point, perhaps the best that can be hoped for is a general index of an individual's repertoire of cognitive strategies. Measurement models appropriate for the assessment of cognitive strategies are most likely to be from the class which assumes an underlying continuous variable. Though it may be possible in the future to identify specific cognitive strategies which are acquired in an all-or-none fashion, at present the continuous true score model appears more manageable and interpretable.

#### Information

Information refers to names attached to objects or to concepts, and to facts or stated relationships between objects or concepts. The acquisition and retention of information is necessary for communication, for facilitating the learning of other types of capabilities, and, very possibly, for any sort of conscious thought above a superficial level.

Gagné and Briggs (1974) discuss three types of information: labels, facts, and bodies of knowledge. Labels are simply names attached to objects or concepts. Evidence that a label has been acquired is shown when an individual can respond to a particular object or example of a concept by stating its name. It is important to emphasize the difference between naming a concept and acquiring the intellectual

skill implied in being able to use the concept. Information has been acquired if an example of the concept elicits a name. Acquisition of an intellectual skill does not require that the concept be named (although the name is usually known). Rather, any example of the general class of objects, events, or relationships which is defined by the concept must be recognized as a member of the class.

Facts are stated relationships between two or more objects or events. Like labels, facts may stand alone. If an individual can state the relationship between given objects or events, then evidence that a fact has been acquired is provided. Acquisition of a fact does not imply the ability to generalize relationships to objects or events not initially presented during the learning of the fact. The ability to generalize would only be expected to occur if an intellectual skill had been acquired.

When interrelated labels and facts are considered as a group, the collection is usually known as a body of knowledge. Bodies of knowledge represent the most common implication of the term "information", and are probably more useful than single labels or facts in dealing with practical problems. Evidence for the acquisition of bodies of knowledge presents problems in logistics and inference. It is rarely feasible to ask individuals to state all of the labels and facts that go into a body of knowledge. Instead, the individual is normally presented with a sample of some of the labels and facts, and if acquisition of the sample is shown, he or she is assumed to have acquired the entire body of knowledge.

Graham and Bergquist (1975) address the problem of choosing an appropriate measurement model for assessing the acquisition of information.

One might argue that single units of verbal information such as labels or single facts are recalled in an all or none manner. Even if this is true, the measurement of single units of information is probably a trivial operation in most instances. Seldom is a single unit of information considered of sufficient importance to be tested separately. More commonly, a collection of information, preferably interrelated to comprise a body of organized knowledge, is tested simultaneously. A collection of information forms a content domain from which items are randomly sampled. Performance of an examinee relative to the entire domain depends upon the number of discrete units of information that have been acquired and remembered. If it is assumed that achievement of each of the discrete units of information is demonstrated independently, any proficiency from 0-100% might be demonstrated on a test. Thus, achievement of verbal information measured by a domain-referenced test would be demonstrated as a continuous variable. (p.3)

#### Attitudes

Attitude is a term used to characterize the internal conditions which affect an individual's behavior towards the external environment. Attitudes may refer to a system of beliefs, to an internal condition arising as a consequence of a conflict in beliefs, or to feelings or emotions. Gagné and Briggs (1974) suggest a more behavioristic point of view. They define attitude as "an internal state which affects an individual's choice of action toward some object, person, or event" (p.62). This definition provides a rationale for the assessment of attitudes.

Choices of action are observable. If it can be assumed that certain choices occur only if an attitude has been acquired, then it is reasonable to assess attitude acquisition by means of observing the

action an individual takes in a choice situation. Gagné and Briggs suggest that the acquisition of attitudes be expressed as the proportion of times a particular action is taken in a given test situation, or as the probability that one action will be chosen over another. The class of measurement models appropriate for attitude assessment of this type is that assuming an underlying continuous variable.

#### Motor Skills

Motor skills are the capabilities required for smooth and purposeful muscular-skeletal movement. Merrill (1971) discusses three categories of motor skills: single responses, response chains, and skilled performances. A single response occurs when a single muscular-skeletal reaction is elicited in the presence of a particular stimulus. Evidence for the acquisition of a single response is shown in three ways. The first is reliability. Reliability implies that the desired response, rather than some other response, occurs whenever the appropriate stimulus is presented, and that it does not occur in the presence of an inappropriate stimulus. Acquired single responses are also characterized by a relatively short latency period between the stimulus presentation and the response, and by their voluntary initiation but involuntary execution.

Examples of single responses are very rare in practical situations. A more realistic level of motor skills is the response chain. Response chains consist of a series of coordinated single responses which represent a single complete performance. An example is swinging a baseball bat. The performance of interest includes the smooth initiation of the swing and continues through the follow through. While swinging a bat



could be analyzed in terms of the multitude of individual single responses, for most purposes the overall performance of the swing is of greatest importance. Evidence for the acquisition of response chains is similar to that for single responses. Once a response chain has been adequately acquired, it is characterized by its reliability, short latency, and the smooth involuntary occurrence of the series of single responses following voluntary initiation of the chain.

Skilled performance requires the coordination of several response chains in the presence of a set of stimuli. Skilled performances are complex and difficult capabilities. They require that each component response chain be fully acquired. They also require that the individual be able to distinguish between a variety of stimuli and be able to respond with the appropriate response chains. Gagné and Briggs (1974) refer to Fitts and Posner (1967) in discussing how such skilled performances come about. In addition to the acquisition of each component response chain, Fitts and Posner hypothesize an executive sub-routine, which is the internal cognitive thought processes which coordinate the skilled performance. The learning of a skilled performance therefore requires the development of an executive sub-routine in addition to the learning of the required muscular-skeletal performances.

Assessing skilled performance presents difficult problems. Individuals vary with respect to the degree of precision with which they can carry out the performance, and the speed at which they can perform. While absolute limits may exist that characterize the best performance that can be achieved, these limits are generally not known, and are probably not very important. For example, at one time running a

four-minute mile was the best that could be expected of anyone. Four-minute miles will no longer win track meets, and it is presumptuous to hypothesize how fast a mile will eventually be run.

It seems to be illogical to discuss whether a skilled performance has been acquired in absolute terms. Rather, the performance must be described in terms of whether it is adequate relative to some standard. The standard will vary from situation to situation. For example, the standards for running a mile for an athlete in condition will be quite different from those for an individual trying to maintain good health. In such cases, the goals of the individual help dictate the standards. In other situations, where individuals act as part of a group, the system may dictate the standards. Setting standards within the context of a system is discussed by Glaser and Klaus (1963), "In practice, proficiency standards can be established at any value between the point where the system will not perform at all and the point where any further contribution from the human component will not yield any increase in system performance" (p.424).

Choosing an appropriate measurement model for motor skills presents many of the same problems as those for information. If single responses or response chains are to be measured, dichotomous true score models appear to be most appropriate. For the assessment of skilled performance, continuous true score models seem to be required.

#### Measurement Models

Six alternative measurement models will be discussed in this review of literature. They were chosen partly on the basis of their availability in the literature, and partly to represent a wide variety

of approaches that may help solve the criterion-referenced measurement problem. Two of the models assume that true score is a dichotomous variable. The other four assume a continuous true score. All six models assume that responses are scored dichotomously and that responses are locally independent for a given individual. In other words, an individual can only get an item right or wrong (as opposed to being able to get partial credit) and responses to any given item are not dependent on responses to any other item.

The dichotomous true score models were developed by Emrick and Adams (1970) and Macready and Dayton (1975). The continuous true score models were developed by Kriewall (1969, 1972) and Millman (1972), Lord and Novick (1968), Novick and Lewis (1974), and Rasch (Wright and Panchapakasan, 1969). The following section of this review treats each model in detail. A more complete discussion of criterion-referenced measurement models can be found in Millman (1974), Hambleton, Swaminathan, Algina, & Coulson (1978), and Steinheiser, Epstein, Mirabella, & Macready (1978).

#### The Emrick and Adams Model

Emrick and Adams (1970) and Emrick (1971a, 1971b) have developed an evaluation model for mastery testing based on the assumption that objectives can be derived which reflect unitary and explicitly defined skills. The model assumes that mastery for each skill is an all-or-none variable. Appropriate tests of skill mastery consist of items which are highly homogeneous in terms of content, form, and difficulty. For such tests, the model assumes that each item provides an unbiased estimate of an individual's mastery status with respect to the skill being measured.

Two types of measurement error are associated with items that fit the model. A false positive error occurs when an individual whose true status is nonmaster ( $\bar{M}$ ) answers an item correctly (e.g., lucky guesses). A false negative error occurs when an individual whose true status is master ( $M$ ) answers an item incorrectly (e.g., careless error). These relationships between true mastery state and measurement error are represented in Table 1.

Expected score distributions for masters and nonmasters follow the familiar binomial distribution:

$$P(c, w|M) = \binom{n}{c} (1-b)^c b^w, \text{ and} \quad (1)$$

$$P(c, w|\bar{M}) = \binom{n}{c} a^c (1-a)^w, \quad (2)$$

where,  $n$  is the number of items on the test,  $c$  is the number of correct responses,  $w$  is the number of incorrect responses,  $a$  is the probability of a correct response from a nonmaster,  $b$  is the probability of an incorrect response from a master, and  $\binom{n}{c}$  is the binomial coefficient for  $c$  successes in  $n$  trials. The expected distribution of correct and incorrect responses for the overall group of examinees is then,

$$\begin{aligned} P(c, w) &= P(M)P(c, w|M) + P(\bar{M})P(c, w|\bar{M}) \\ &= P(M) \binom{n}{c} (1-b)^c b^w + P(\bar{M}) \binom{n}{c} a^c (1-a)^w. \end{aligned} \quad (3)$$

Table 1 shows the relationship between true mastery state, observed responses to a single item, the probability of a false positive error, and the probability of a false negative error. The probabilities of false positive and false negative errors are treated as response contingencies and a phi coefficient is computed, indicating the correlation between observed responses on a single item and true mastery state (Emrick, 1971a, p.323):

		Observed Response	
		Wrong	Correct
True Mastery	Master	b	1-b
State	Nonmaster	1-a	a

**Table 1: True Mastery State and Measurement Error for the Emrick Model**

a = the probability of a correct response from a nonmaster

b = the probability of an incorrect response from a master

1-a = the probability of a valid nonmaster incorrect response

1-b = the probability of a valid master correct response

$$\phi = (1-a-b)/\sqrt{1-(a-b)^2} \quad (4)$$

A second estimate of the correlation between observed responses and true mastery state is obtained by computing the average interitem correlation among the items on the test. Average interitem correlation was estimated by Emrick by computing the test reliability using the Kuder-Richardson formula 20 and then adjusting the reliability to that of a single item using the Spearman-Brown prophecy formula. Since reliability is defined as the proportion of total variance that is true variance, it can be interpreted as an unbiased estimate of the squared correlation between an examinee's true mastery state and his or her item response.

By equating item reliability with  $\phi$  (squared), item responses, true mastery state, and error probabilities are directly related. If the ratio of the probabilities of false positive to false negative errors is known (or if it can be estimated), values for the probabilities can be calculated.

Epstein (1978) and Wilcox and Harris (1977) have shown that the analysis as described in the model is only appropriate if the proportion of masters equals the proportion of nonmasters in the group of examinees. The correct relationship between the reliability of a single item, the measurement errors, and the proportions of masters and nonmasters is

$$r_1 = (\phi)^2 = \frac{[1-a-b]^2}{[1-a-b+2ab+P(N)(b-b^2)/P(\bar{N})+P(\bar{N})(a-a^2)/P(N)]} \quad (5)$$

where  $r_1$  is the reliability of a single item (Epstein, 1978, p.51; Wilcox and Harris, 1977, p.217). For the case where  $P(M)=P(\bar{M})$ , the above equation reduces to the form described by Emrick in equation (4),

$$r_1 = (\phi)^2 = (1-a-b)^2 / (1-(a-b)^2).$$

In order to operationalize the model, the test developer must provide estimates for the ratios of the probabilities of false positive to false negative errors and  $P(M)$  to  $P(\bar{M})$ , and  $r_1$  must be calculated. Equation (5) (or equation (4), if appropriate) can then be solved. Although the estimates are subjective, a logical analysis of the testing situation combined with experience in using the model should lead to realistic values. For example, the ratio of the probabilities of false positive to false negative errors for a four response multiple choice test is likely to be much greater than the ratio of the probabilities of false positive to false negative errors for a constructed response test. This is simply because chance alone allows nonmasters to get some items correct on the multiple choice test, while the likelihood of a nonmaster guessing the correct response to a constructed response item is relatively low. Similarly, the ratio of masters to nonmasters in the examinee population should not cause undue problems, particularly if the instruction has been well designed and systematic steps have been taken to control student learning. For example, if results from a post-test are being analyzed, the ratio of  $P(M)$  to  $P(\bar{M})$  should be relatively high. In a pre-test situation the opposite would be the case. A conservative estimate for the ratio of  $P(M)$  to  $P(\bar{M})$  is 1.0, and may prove useful as a starting point until more experience is gained in using the model.

Emrick proposed that mastery cutoff scores be optimized in terms of the relative costs of incorrect mastery/nonmastery decisions, and the previously determined parameters. The optimization formula is

$$k = \frac{\log [b/(1-a)] + (1/n) \log [L_2 P(M)/L_1 P(\bar{M})]}{\log [ab/\{(1-a)(1-b)\}]} \quad (6)$$

where,  $k$  is the percent of items correct required for a mastery decision,  $a$  is the probability of a false correct response,  $b$  is the probability of a false incorrect response,  $L_1$  is the cost associated with a false pass decision,  $L_2$  is the cost associated with a false fail decision,  $n$  is the number of test items,  $P(M)$  is the proportion of masters in the examinee group, and  $P(\bar{M})$  is the proportion of nonmasters in the examinee group (Emrick, 1971a, p.324).

Emrick (1971b) discusses an empirical validation of the evaluation model for mastery testing. An experiment was conducted in which 96 third grade students were taught to identify three increasingly complex concepts. Tests designed to show their ability to identify members of a group of objects which belong to the concept group were administered following the training. The results were analyzed according to the model. Results for 5 and 10 item forms of 2-option and 4-option multiple choice tests were analyzed. The results from common forms of the post-test were then aggregated and compared to the results of the last 10 training trials. Emrick concluded that, "the evidence derived in support of this model, although not striking or dramatic is nonetheless favorable" (p.49). Because of the problems associated with prior estimation of the proportions of masters and nonmasters (not addressed in the Emrick paper), the small sample size and complexity of the experimental design, and a rather confusing discussion of the model



validation procedure, further research should be conducted before any conclusions are reached concerning the appropriateness of the model.

#### The Macready and Dayton Model

The Macready and Dayton model (1975) is a special case of a general probabilistic model developed by Dayton and Macready (1976) for validating behavioral hierarchies. The general model provides great flexibility by allowing for a wide variety of true response patterns and by allowing measurement error to be item specific. The cost of this flexibility is that for the more complex models a relatively large number of test items is required and, for all models, a large subject population is required to obtain stable parameter estimates.

Macready and Dayton argue that a reasonable approximation of the more general model for criterion-referenced testing purposes is obtained if mastery is defined as an all-or-none variable. Under this assumption, the only allowable true response patterns would be all correct or all incorrect. Measurement errors are (1) the probability of a non-master guessing the answer to an item correctly, and (2) the probability of a master missing an item. If the probabilities of each type of error are constant for all items on a given test, then the Macready and Dayton model begins with the same statistical model as the Emrick and Adams model. Macready and Dayton also allow for the more complex case where the error probabilities are item specific.

For example, if a four item test were given, the assumption that the true score be an all-or-none variable requires that the only error free response patterns are (0,0,0,0) for nonmasters, and (1,1,1,1) for masters. For the general case, the probabilities of a nonmaster passing items are  $a_1$ ,  $a_2$ ,  $a_3$ , and  $a_4$  for each item respectively.

Similarly, the probabilities of a master getting items incorrect are  $b_1, b_2, b_3$ , and  $b_4$ . In general, there will be  $2^n$  possible response patterns for an  $n$  item test. The necessary equations for the probability of response pattern  $j$  occurring under the general model are

$$P(j|M) = \prod_{i=1}^n b_i^{(1-x_{ij})} (1-b_i)^{x_{ij}} \quad (7)$$

for masters, and

$$P(j|\bar{M}) = \prod_{i=1}^n a_i^{x_{ij}} (1-a_i)^{(1-x_{ij})} \quad (8)$$

for nonmasters, where  $i$  is the item number from 1 to  $n$ , and  $x_{ij}$ , which can equal 0 or 1, is the score on the  $i$ th item for response pattern  $j$  (p.3). When the equations for masters and nonmasters are combined, the probability of the  $j$ th response pattern is

$$P(j) = P(M)P(j|M) + P(\bar{M})P(j|\bar{M}). \quad (9)$$

For example, for response pattern (0,1,1,0), the necessary equations are as follows. For masters, the probability of the above response pattern is  $b_1 \times (1-b_2) \times (1-b_3) \times b_4$ . The four terms are necessary to account for the different measurement errors for different items. For nonmasters, the probability of the response pattern above is  $(1-a_1) \times a_2 \times a_3 \times (1-a_4)$ . Combining these results, the probability of observing the above response pattern is

$$P(0,1,1,0) = P(M)b_1(1-b_2)(1-b_3)b_4 + P(\bar{M})(1-a_1)a_2a_3(1-a_4). \quad (10)$$

For the simpler case of equal probabilities of error across test items, the equation reduces to

$$P(0,1,1,0) = P(M)b^2(1-b)^2 + P(\bar{M})a^2(1-a)^2. \quad (11)$$

One more important difference between the two forms of the model should be noted. For the general form of the model, the response pattern is required to calculate the probabilities of interest. That

is,  $P(0,1,1,0) = P(M)b_1(1-b_2)(1-b_3)b_4 + P(\bar{M})(1-a_1)a_2a_3(1-a_4)$  is not equal to  $P(1,0,0,1) = P(M)(1-b_1)b_2b_3(1-b_4) + P(\bar{M})a_1(1-a_2)(1-a_3)a_4$  even though both response patterns indicate two correct responses. However, in the simpler case of the model the probabilities are equal,  $P(0,1,1,0) = P(1,0,0,1) = P(M)b^2(1-b)^2 + P(\bar{M})a^2(1-a)^2$ , and only the number correct is required for carrying out the calculations. Since the binomial coefficient indicates the number of ways a given number of successes can occur in  $n$  trials, the final result under the simpler case is the same as the Emrick and Adams basic equation (equation (3)),

$$P(c,w) = P(M) \binom{n}{c} (1-b)^c b^w + P(\bar{M}) \binom{n}{c} a^c (1-a)^w.$$

The general case of the Macready and Dayton model requires that  $2n + 1$  parameters be estimated for an  $n$  item test. The  $2n + 1$  figure is obtained from the probabilities of false positive and false negative errors for each of the  $n$  items, plus either  $P(M)$  or  $P(\bar{M})$  since  $P(M) + P(\bar{M}) = 1$ . For the simpler case, only three parameters,  $a, b$ , and  $P(M)$  or  $P(\bar{M})$  must be estimated. Macready and Dayton obtain the parameter estimates by using maximum likelihood procedures. It is beyond the scope of this presentation to go into their procedure in detail. However, it should be noted that the procedure, in general, attempts to find values for the necessary parameters that will closely reproduce the observed data. It also provides estimates of the variance of the parameter estimates, which may prove useful in evaluating the acceptability of the model in specific instances.

Once the parameters have been estimated the model can be used for decision making. The procedure is logical and straightforward. For the general case, the probabilities of masters and nonmasters obtaining the

j response patterns are calculated. Mastery versus nonmastery classification rules are then established for each response pattern. For the simpler case, the probabilities of masters and nonmasters obtaining zero through n items correct are calculated, and a cutoff score is chosen such that the probability that a master would achieve a score below the cutoff score plus the probability that a nonmaster would achieve a score at or above the cutoff score is minimized. In both cases the strategy is to minimize the total misclassification for the examinee group. Macready and Dayton have computer programs available for analyzing data. In addition, they provide tables showing optimal cutoff scores for various test lengths, parameter estimates and loss ratios in their 1975 paper.

#### The Proportion Correct Model

The first and least complex of the models which assume that mastery is a continuous variable is based on the proportion of items answered correctly on an n item test. The basic model has been developed theoretically and operationalized by Kriewall (1968, 1972). Millman (1972) discussed the model's practical applications and developed useful and easy to understand tables which may be used in test development. A unique aspect of the model is that the procedures and their applications to real problems are independent of sample data. The other models discussed here, and, in fact, most psychometric models, use observed examinee data to estimate parameters. The sample free nature of the Kriewall and Millman approach is very appealing for criterion-referenced testing. It is the only method that does not compare examinees in estimating abilities. Since the other models use observed scores to

estimate parameters, examinees are indirectly being compared to one another.

The model assumes that the items on a test are a random sample of items from a well defined domain. The domain must be sufficiently homogeneous for all the items within the domain to be equally difficult for a given individual. This does not mean that items will have equal difficulty in traditional psychometric terms. More capable individuals will find the items easier than less capable individuals. However, for any given examinee, the probability that he or she will respond to an item correctly is the same for all items within the domain. Kriewall defines "proficiency" as the probability of a correct response. It may vary from 0 to 1.0, and will be denoted  $p$ .

The model also assumes that items are locally independent. Independence of items is not an obvious concept. Local independence of items implies that a person's response to any given item on the test is statistically independent of his response to any other item.

To state it another way, in an infinite subpopulation of examinees, all of whom are at the same ability level, scores on one test item will be statistically independent of scores on another. It will be recognized that the assumption of local independence does not imply that test items are uncorrelated over the total group of examinees (Lord and Novick, 1968, p.361). Correlations between items measuring the same ability will, in general, exist whenever the examinees responding to the items differ on the underlying ability measured by the test. (Hambleton and Traub, 1973, p.196)

The basic equation for the model is the distribution of the number correct score for a given proficiency over repeated random samples of  $n$  items from the domain. It is binomial with parameter  $p$ , the proficiency:

$$f(x|p) = \binom{n}{x} p^x(1-p)^{n-x} \quad (12)$$

where,  $x$  is the number of items answered correctly given proficiency  $p$ .

The error of measurement for a given individual with proficiency  $p$ , expressed in terms of the number of items erroneously missed or passed, will be denoted  $e_p$ .  $e_p$  for an  $n$  item test is

$$e_p = x - np. \quad (13)$$

Since the expected value of  $x$  for an  $n$  item test is  $np$ , the expected value of the error of measurement, for repeated testing, is zero. More specifically, it can be shown (Lord and Novick, 1968, p.458) that the obtained proportion correct,  $x/n$ , is the maximum likelihood estimator of the true proportion correct or the proficiency. The estimate has a variance of  $p(1-p)/n$  which can be made as small as desired by sufficiently increasing  $n$ . This implies that longer tests provide better estimates of proficiency than shorter tests.

The model provides probabilistic information about test performance for any ability of interest. In some cases this information will be all that is required by the decision maker. More frequently, an easy to use rule for categorizing students will be desired. Such a rule can be developed according to the following scheme.

Two abilities must be identified, a minimum mastery proficiency, and a maximum nonmastery proficiency. Minimum mastery represents the proficiency an individual must have with respect to the domain to be considered a master. Rarely will 100% mastery be required. More common levels of minimum mastery might be 70%, 80%, or 90%, depending on the importance of the material and the level of competency desired. It is important to keep in mind that proficiency may be defined as the

proportion of all possible test items in the domain that would be answered correctly. It may also be defined as the probability of responding correctly to a randomly chosen item from the domain. It is not necessarily the percent of the items that must be answered correctly on a particular test in order to pass.

The maximum nonmastery proficiency is the highest level of proficiency an individual could attain over the domain and yet not be considered a master of the material. Maximum nonmastery levels are often about 50%. That is, it is often considered reasonable to assume that even if an examinee knew half of the material included in the domain, he could not be considered capable enough to be called a master. Any proficiency between the minimum mastery proficiency and the maximum nonmastery proficiency falls within an indifference region. That is, it makes no practical difference whether an individual with a proficiency that falls within the indifference region is classified a master or a nonmaster. In general, the larger the indifference region, the smaller the number of test items required for decision making.

The probabilities for achieving any given score on an  $n$  item test for minimum masters and maximum nonmasters can be calculated by applying the basic equation. Misclassification occurs when nonmasters are classified as masters and masters are classified as nonmasters. The desired decision rule is one which has a cutoff score which minimizes the probability of misclassification. The probability of misclassification for masters as nonmasters is

$$b = \sum_{x=0}^{c-1} \binom{n}{x} p_m^x (1-p_m)^{n-x}, \quad (14)$$

and the probability of misclassifying a nonmaster as a master is

$$a = \sum_{x=c}^n \binom{n}{x} p_m^x (1-p_m)^{n-x}, \quad (15)$$

where,  $n$  is the number of test items,  $x$  is the number of correct responses,  $c$  is the minimum number of correct responses required for a mastery decision or the cutoff score,  $p_m$  is the minimum mastery proficiency, and  $p_{\bar{m}}$  is the maximum nonmastery proficiency. By carrying out the above calculations with various values of  $c$ , an optimal cutoff score can be determined along with its probabilities of misclassification.

It is important to realize that the model represents only a gross approximation of reality. The model deals with only two proficiency states explicitly. This would be fine if all examinees had proficiencies equal to the minimum mastery or maximum nonmastery proficiencies. Of course this will never be the case. Fortunately, the model is conservative. The probabilities of misclassification for examinees with proficiency above minimum mastery or below maximum nonmastery must be less than the probabilities of misclassification for examinees at these levels. Since examinees with proficiencies in the indifference region are no problem, the actual number of misclassifications is expected to be less than that predicted by the model.

#### The Binomial Error Model

A natural extension of the proportion correct model is the binomial error model (Lord and Novick, 1968). The binomial error model is more powerful than the simple proportion correct model because the entire distribution of observed responses is included in the analysis. All of the assumptions discussed with respect to the proportion correct



model hold for the binomial error model. The conditional distribution for observed score  $x$  for given true proportion correct  $p$  is the binomial

$$h(x|p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad (16)$$

where,  $n$  is the total number of items on the test. It is also assumed that items are scored dichotomously, that total score for an examinee is the number of items answered correctly, that items are locally independent, and that items are equally difficult for a given examinee.

An addition to the proportion correct model is the specification of the relationship between the observed score distribution and the underlying true proficiency distribution

$$\phi(x) = \binom{n}{x} \int_0^1 g(p) p^x (1-p)^{n-x} dp, \quad (17)$$

where  $\phi(x)$  is the distribution of the observed scores, and  $g(p)$  the unknown distribution of true scores (Lord and Novick, 1968, p.512).

Lord and Novick (1968) show that if the regression of true score on observed score is linear then the distribution of the observed scores for the entire examinee group, symbolized  $h(x)$  to distinguish this special case from the general case  $\phi(x)$ , is negative hypergeometric

$$h(x) \equiv \left[ \frac{s^{(n)}}{(r+s)^{(n)}} \right] \left[ \frac{(-n)_x (r)_x}{(-s)_x x!} \right], \quad (18)$$

where  $r$  and  $s$  are parameters to be determined,  $s^{(n)}$  is defined as  $s(s-1)\dots(s-n+1)$ ,  $(s)_x$  is defined as  $s(s+1)\dots(s+x-1)$ , and  $s^{(0)}$  and  $(s)_0$  are defined to equal 1 (Lord and Novick, 1948, p.516). The parameters,  $r$  and  $s$ , can be expressed in terms of the moments of the observed score distribution as follows (Lord and Novick, 1968, p.517):

$$r = (-1 + 1/\alpha_{21})\mu_x, \quad (19)$$

$$s = -r - 1 + n/\alpha_{21}, \text{ and} \quad (20)$$

$$\alpha_{21} = [n/(n-1)] [1 - \mu_x (n - \mu_x) / n\sigma_x^2]. \quad (21)$$

Lord and Novick (1968) prove a very useful consequence of the model. "Under the binomial error model, if the observed score distribution is negative hypergeometric, then the regression of true score on observed score is linear" (p.517).

The discussion thus far has outlined an internal check of the appropriateness of this model for any given data set. That is, if one can show adequate fit to the negative hypergeometric distribution by the observed scores, then it is reasonable to continue with this model assuming linear regression. If adequate fit is not obtained, then the more general nonlinear regression approach must be used, or alternative models must be identified.

Lord and Novick (1968) show that if the observed score distribution is negative hypergeometric, the true score distribution is either the two parameter beta distribution, or some other distribution having identical moments up through order  $n$ . In either case, they show (p.521) that the regression of true proficiency on observed score is given by the linear equation

$$E(p|x) = \frac{\alpha_{21}}{n} x + \frac{(1 - \alpha_{21})\mu_x}{n}. \quad (22)$$

Epstein (1975) provides an example of the use of the binomial error model for criterion-referenced testing. The data described in his paper were shown not to statistically significantly deviate from the negative hypergeometric distribution using a chi-square goodness of

fit test, and the appropriate regression equation was calculated. He then calculated true score estimates for each observed score and suggested that the obtained true score estimates be used for decision making instead of the raw observed proportion correct scores. Epstein did not specify a particular decision making model in his paper. However, the situation described clearly fits the Roudabush Case IV situation; a pseudo continuous measure of a continuous true score, described earlier.

#### The Beta-Binomial Bayesian Model

The binomial error model builds on the simple proportion correct model by using group data and an assumption concerning the form of the underlying true score distribution in computing true score estimates from observed scores. Novick and Lewis (1974) introduce information which is known about the performance of examinees before testing as a prior distribution in their development of a Bayesian procedure for criterion-referenced decision making.

A reasonable choice for a prior distribution is one that is a member of the Beta family of distributions (Novick and Jackson, 1974). Recall that one of the theoretical results of the binomial error model is that if the observed score distribution fits a negative hypergeometric distribution as required, then the true score distribution will be a member of the Beta family. Beta distributions can take on a variety of forms including a uniform distribution of proficiency from 0 to 1.0, a close approximation to the normal distribution, a U shaped distribution, and extremely skewed distributions in either direction. For the case of dichotomously scored tests where the conditional (on

proficiency) observed score distribution is binomial, a Beta prior distribution combined with the observed data yields a Beta posterior distribution. In fact, if the prior distribution is  $B(r,s)$  and  $x$  correct responses are observed for  $n$  items, then the posterior distribution is  $B(x+r, n-x+s)$ .

The procedure is extremely easy to use, once a prior distribution has been specified. One simply determines the appropriate posterior distribution for each observed score and then finds the probability that the proficiency equals or exceeds some criterion proficiency. If the probability is sufficiently high, the examinee is classified a master. Otherwise, a nonmaster classification is made. For example, consider a case where little is known about the examinee group. A reasonable choice of a prior distribution is that proficiency is uniformly distributed,  $B(1,1)$ . If a examinee score of 7 correct on a 10 item test is observed, then the posterior distribution is  $B(7+1, 10-7+1) = B(8,4)$ . The probabilities that the examinee's proficiency is greater than or equal to .60, .70, and .80 are .88, .69, and .38, respectively. If the criterion proficiency had been set at .70 and a probability of .5 or better had been set for a mastery decision, then such a student would be classified a master. If, however, the criterion proficiency was .80, then using the .5 or better decision rule, the student would be classified a nonmaster.

Novick, Lewis, and Jackson (1973) discuss methods for determining the parameters of the prior distribution, Novick (1973) describes the Computer Assisted Data Analysis (CADA) system which guides a decision maker through the process, and the Novick and Lewis (1974) article

contains tables suggesting prior distributions for a series of instructionally relevant situations with the appropriate posterior distributions and probabilities for a variety of test lengths and observed scores.

#### The Rasch Model

A relatively new approach to psychological measurement is based on the Rasch logistic model (Wright, 1968; Wright and Panchapakesan, 1969; Whitely and Dawis, 1974). According to Wright, "The model says simply that the outcome of the encounter (between an individual and a test item) is governed by the product of the ability of the person and the easiness of the item" (p.88). If this claim is true, it would seem that the Rasch model represents an ideal tool for criterion-referenced testing. Yet, as Whitely and Dawis (1974) point out, "To date, however, the Rasch model has had little apparent impact on test development. The reasons for this are not clear, particularly since initial research has been encouraging" (p.164).

Tests which fit the Rasch model have the following specific properties: (1) the estimated values of the item easiness parameters will not vary significantly over different samples of people, (2) the estimate of a person's ability, given a raw score, will be invariant over different samples of people, and (3) the estimates of a person's ability from any subset of Rasch calibrated items will be statistically equivalent.

In order for the Rasch model to be applicable, several basic assumptions must be met. The first assumption is that subjects and items are locally independent.

Independence of subjects means that the item responses of any given person do not affect the responses of any other person. Independence of items, on the other hand, means that a person's responses to preceding items do not affect his responses to later items. Thus, the probabilities a person will pass the various individual items must remain invariant, regardless if the ability test contains the whole item pool or only some subset of items. (Whitely and Dawis, 1974, p.165)

Items comprising a test which fits the Rasch model are assumed to all be measuring a single unidimensional latent trait. What this means, practically, is that the items must be homogeneous in the sense that they all measure the same single ability. Statistically, unidimensionality implies "that if subjects are grouped according to raw score, within each group, there will be no remaining significant correlations between items. Thus, all of the covariation between the items (over the total group of examinees) is accounted for by variation of persons on the latent trait (ability) to be measured" (Whitely and Dawis, 1974, p.165).

Discrimination refers to the quality of an item in terms of the information it provides about levels of ability. For example, if all individuals, regardless of ability, passed an item, that item would have a discrimination of zero. It provides no information about level of ability. Discrimination is a function of the rate at which the probability of passing an item increases with increasing ability. Items which fit the Rasch model are assumed to have equal discrimination. The Rasch model does not contain a parameter associated with discrimination. It should be noted that equal discrimination does not imply anything about item easiness. Clearly, a range of easiness is required for practical testing. Items can be equally discriminating

for different ranges of the ability continuum, and therefore be unequal in easiness. The final assumptions of the Rasch model are that guessing is negligible and that there are not errors in scoring.

The mathematical properties of the model can be most easily described in terms of an item by total raw score group matrix. For an  $n$  item test, such a matrix will have  $n$  rows, one for each item, and  $n-1$  columns, one for each total raw score,  $1, 2, \dots, n-1$ , except 0 and  $n$  correct. Total raw scores of 0 and  $n$  correct are excluded because they provide no information about the items. Each cell represents the probability,  $P_{ij}$ , that an individual with ability  $A_j$  will pass item  $i$  with easiness parameter  $E_i$ . The Rasch probability function (Whitely and Davis, 1974, p.164) is

$$P_{ij} = (A_j \times E_i) / (1 + A_j \times E_i). \quad (23)$$

In order to estimate the Rasch item and person parameters, the cell probabilities must be converted to likelihood ratios. Likelihood ratios are most easily thought of as betting odds and are defined as the ratio of the probability of passing to the probability of failing:

$$\begin{aligned} \text{Likelihood} = P_{ij} / (1 - P_{ij}) &= \frac{(A_j \times E_i) / (1 + A_j \times E_i)}{1 - (A_j \times E_i) / (1 + A_j \times E_i)} \\ &= \frac{(A_j \times E_i) / (1 + A_j \times E_i)}{1 / (1 + A_j \times E_i)} \\ &= A_j \times E_i. \end{aligned} \quad (24)$$

Converting to logarithms allows for simpler computations and shows that, on a logarithmic scale, the log-likelihood that a person will pass an item is simply the sum of the log of his ability and the log of the easiness of the item. Symbolically, these relationships are

indicated as follows:

$$t_{ij} = \log P_{ij} / (1 - P_{ij}), \quad (25)$$

$$b_j = \log A_j, \quad (26)$$

$$d_i = \log E_i, \quad (27)$$

and, from equation 24 above,

$$t_{ij} = b_j + d_i. \quad (28)$$

Wright and Panchapakesan (1969) and Wright and Mead (1975) have published computer programs to estimate Rasch parameters using maximum likelihood procedures. For each item, its easiness parameter estimate and the standard error of the estimate is calculated. Similarly, for each raw score group, its ability parameter estimate and the standard error of the estimate are calculated. Goodness-of-fit information is calculated and a variety of descriptive statistics, tables, and graphs are provided.

Kifer and Bramble (1974) describe an attempt to use the Rasch model to calibrate a criterion-referenced test. They also discuss how Rasch model ability estimates can be interpreted in terms of criterion-referenced testing. The general procedure they followed consisted of (1) an initial attempt to calibrate the item pool, (2) based on the results of the initial calibration, elimination of items which did not fit the model, (3) recalibration of the item pool, and (4) estimation of abilities. The interpretation problem was to determine whether a particular score exceeded some criterion required for mastery. Kifer and Bramble's procedure follows.



Given any criterion, we assume that the estimate of latent ability at the criterion is an estimate of the 'true' ability at that point. The standard error of measurement associated with that ability level is assumed to be an estimate of the observed score distribution around the 'true' criterion. Based on these assumptions, it is possible to ask the question of the probability that any observed score comes from that particular distribution. Although the choice of sampling distribution for our estimates is arbitrary, because maximum likelihood estimates are asymptotically normal, we choose the normal distribution.  
(p.4)

The probability information available from such an interpretation of criterion and obtained scores can be used to estimate the probabilities of misclassifying masters and nonmasters. This information, along with the costs associated with misclassification, forms the basis for decision making.

The Rasch model, as implemented for criterion-referenced testing, clearly falls into the category of continuous true score models. It seems to offer great potential for supplying ability estimates which can be interpreted in absolute terms. It also seems to offer considerable flexibility in designing decision making procedures. A final implication of the model lies in the interpretation of the criterion scores. One problem with most criterion scores expressed as a percentage of the domain that must be mastered is interpreting a statement such as 80% capable. The usual interpretation is that an individual with 80% capability is expected to be able to do 80% of the items in the domain. "Which 80%", is never answered. The latent trait theory underlying the Rasch model may help. Rather than 80%, one may determine a criterion ability. Since ability is invariant from one set of calibrated items to the next, the question of "which 80%" is

irrelevant. Further research is required to substantiate Rasch model claims for criterion-referenced domains of test items, to help in the interpretation of Rasch ability estimates, and to establish procedures for specifying criterion ability.

The Rasch model is only one of a class of measurement models known as latent trait models. Latent trait models and the theory on which they are based are receiving increasing attention in the literature. For example, the summer 1977 issue of the Journal of Educational Measurement is devoted to applications of latent trait models. Latent trait models other than the Rasch model may include unique discrimination parameters for each item, parameters to account for guessing, techniques to utilize all of the information contained in multichotomously scored items, or approaches to deal with multidimensional tests. A thorough review of recent developments is provided by Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978). Despite the attractive features of latent trait theory for criterion-referenced testing, the Hambleton, et al. article points out that, "To date, only a minimal amount of research has been done concerning the applicability of latent trait models to criterion-referenced tests" (p.496).

### 3. METHODS

The general approach taken in this study is based on the contention that the ideal case for investigating criterion-referenced testing and decision making is one in which the true abilities and measurement error free test scores of the participating individuals are known. Clearly, for empirical research, this is an impossible goal. However, a reasonable approximation of a score free of measurement error may be obtained if a very large number of items can be sampled from a domain of interest and included on a test. The obtained approximate true score can then serve as a criterion score for investigations of the characteristics of tests of more realistic numbers of items. Given this general approach and the objectives of this study, the primary methodological considerations are those relating to choosing a suitable data base, describing the test characteristics, choosing and implementing the measurement models, and comparing the models.

A variety of data analyses are described in this and the results section of this study. All analyses requiring computer assistance were conducted using a Department of the Army UNIVAC 1108 computer located in Edgewood, Maryland. With the exception of several procedures using the Statistical Package for the Social Sciences (Nie, Hull, Jenkins, Steinbrenner, and Bent, 1975), all other programming was done by the author.

### The Data Base

The data for this study are .45 caliber pistol marksmanship scores obtained by military police trainees on the Military Police Firearms Qualification Course (MPFQC) (US Army Military Police School, 1975). The MPFQC is used to certify trainees in pistol marksmanship, is required for graduation from the school, and is administered immediately following training. Under normal circumstances, the test consists of 50 rounds fired from eight stations, called tables by the school, differing in shooting position and distance to the target. The tables were chosen by the school to represent a cross section of the kinds of problems encountered by military police on the job (Figure 1).

The MPFQC represents a suitable data base for this study for several reasons. First, it was designed as a criterion-referenced test. Since the desired behavior is well defined and all items on the test are representative of the behavioral domain of interest, it satisfies the definition of a CRT offered earlier. Second, the behavior required for each shot appears to be equivalent, yielding a homogeneous set of test "items". Third, the test administrators and marksmanship instructors represented a source of expertise that could be called upon for implementing a Bayesian analysis. Finally, and perhaps most importantly, the Military Police School agreed to modify its testing procedure to allow each trainee in this study to shoot a total of 240 rounds.

The 240 rounds were fired in three separate repetitions of 80 rounds each. The first repetition was fired one morning, the second, that afternoon, and the third, the following morning. Each 80 round repetition consisted of firing 10 shots at each of the 8 tables on

Table	Distance	Position
1	35 meters	Prone - Two Hands
2	25 meters	Standing - Two Hands
3	25 meters	Standing - Left Hand
4	25 meters	Standing - Right Hand
5	15 meters	Kneeling - Two Hands
6	15 meters	Kneeling - Left Hand
7	15 meters	Kneeling - Right Hand
8	7 meters	Crouching - Two Hands

Figure 1: The Military Police Firearms Qualification Course

the MPFQC. The 10 shots at each table were further divided into two groups of five shots each. After firing the five shots, the trainees reloaded their weapons, scores were recorded, and the holes in the target were taped to prevent feedback to the trainees. The 240 round score served as the criterion approximate true score for subsequent analyses.

A total of 237 trainees participated in the study. This group represented 10 different classes at the school. The first group of between 20 and 25 trainees in each class to complete their training formed the subject pool for this study. The data were collected from November, 1976 to March, 1977 at the US Army Military Police School.

Analysis of the test results indicated the tables were not homogeneous with respect to difficulty. In fact, the MPFQC clearly consists of two subtests. Tables 1 through 4 are relatively difficult. Tables 5 through 8 are relatively easy. These results influenced the sampling plan for the more realistic subtests and also suggested the need for two additional criterion scores, one based on the 120 hard shots and the other based on the 120 easy shots. Subtests of 10, 20, 40, and 80 shots were sampled according to the following scheme:

- (a) The 10 round subtests were the table scores;
- (b) The 20 round subtests were sampled to produce 6 hard tests (Hard 21 - Hard 26), 6 easy tests (Easy 21 - Easy 26), and 12 tests consisting of both hard and easy tables (Mix 201 - Mix 212);
- (c) The 40 round subtests were sampled to produce 3 hard tests (Hard 41 - Hard 43), 3 Easy tests (Easy 41 - Easy 43), and 6 tests consisting of both hard and easy tables (Mix 41 - Mix 46).
- (d) The 80 round subtests were the repetitions. Descriptive data from all of these subtests and the specific sampling plan can be found in Table 2.

	HARD 21	HARD 24	EASY 21	EASY 24	MIX 201	MIX 202	MIX 207	MIX 208	HARD 41	EASY 41	MIX 41	MIX 44	MEAN	KR21
REP1													.756	.846
TABLE11									X				.673	.654
GROUP1	X				X						X		.673	
GROUP2		X					X					X	.673	
TABLE12									X				.623	.612
GROUP1	X				X						X		.643	
GROUP2		X					X					X	.603	
TABLE13									X				.559	.650
GROUP1	X					X					X		.563	
GROUP2		X						X				X	.555	
TABLE14									X				.654	.647
GROUP1	X					X					X		.688	
GROUP2		X						X				X	.619	
TABLE15			X			X				X			.838	.563
GROUP1				X							X		.837	
GROUP2							X					X	.838	
TABLE16			X		X					X			.824	.620
GROUP1				X							X		.815	
GROUP2						X						X	.833	
TABLE17			X		X					X			.905	.616
GROUP1				X			X				X		.911	
GROUP2												X	.898	
TABLE18			X			X				X			.970	.515
GROUP1				X				X			X		.967	
GROUP2							X					X	.972	
MEAN	.642	.626	.883	.883	.761	.734	.754	.755	.627	.884	.762	.755		
KR21	.649	.776	.661	.682	.627	.566	.624	.711	.823	.744	.733	.793		

Table 2: MPFQC Shot Groups, Tables, and Sampled Subtests;  
Means and Reliabilities  
(an X indicates that shot group or table scores were summed  
to equal subtest scores)

	HARD 22	HARD 25	EASY 22	EASY 25	MIX 203	MIX 204	MIX 209	MIX 210	HARD 42	EASY 42	MIX 42	MIX 45	MEAN	KR21
REP2													.761	.875
TABLE21 GROUP1	X				X				X		X		.683	.633
GROUP2		X					X					X	.693	
													.673	
TABLE22 GROUP1	X					X			X		X		.649	.660
GROUP2		X						X				X	.666	
													.632	
TABLE23 GROUP1	X				X				X		X		.537	.677
GROUP2		X					X					X	.540	
													.534	
TABLE24 GROUP1	X					X			X		X		.670	.712
GROUP2		X						X				X	.673	
													.667	
TABLE25 GROUP1			X		X					X	X		.845	.633
GROUP2				X			X					X	.857	
													.834	
TABLE26 GROUP1			X			X				X	X		.823	.732
GROUP2				X				X				X	.827	
													.817	
TABLE27 GROUP1			X			X				X	X		.906	.744
GROUP2				X				X				X	.909	
													.903	
TABLE28 GROUP1			X		X					X	X		.974	.438
GROUP2				X			X					X	.971	
													.976	
MEAN	.613	.677	.885	.911	.752	.746	.794	.794	.635	.887	.749	.769		
KR21	.764	.739	.600	.615	.631	.590	.697	.585	.843	.796	.748	.767		

Table 2 (cont)



	HARD 23	HARD 26	EASY 23	EASY 26	MIX 205	MIX 206	MIX 211	MIX 212	HARD 43	EASY 43	MIX 43	MIX 46	MEAN KR21
REP3													.791 .855
TABLE31 GROUP1 GROUP2	X	X				X		X	X		X	X	.654 .688 .655 .653
TABLE32 GROUP1 GROUP2	X	X			X		X		X		X	X	.689 .654 .695 .684
TABLE33 GROUP1 GROUP2	X	X				X	X		X		X	X	.620 .679 .620 .619
TABLE34 GROUP1 GROUP2	X	X			X			X	X		X	X	.715 .688 .738 .692
TABLE35 GROUP1 GROUP2			X	X	X			X		X	X	X	.881 .560 .882 .880
TABLE36 GROUP1 GROUP2			X	X	X		X			X	X	X	.865 .662 .862 .869
TABLE37 GROUP1 GROUP2			X	X		X		X		X	X	X	.928 .636 .927 .928
TABLE38 GROUP1 GROUP2			X	X		X	X			X	X	X	.976 .448 .972 .980
MEAN	.643	.662	.891	.914	.765	.769	.788	.788	.670	.913	.767	.763	
KR21	.778	.711	.642	.551	.563	.704	.559	.659	.835	.736	.781	.727	

Table 2 (cont)

### Test Characteristics

The important questions concerning the MPFQC involve the homogeneity of equivalent subtests and the stability and reliability of the test scores. Average scores were computed and operating characteristic curves were plotted to indicate similarities and differences in the subtests. Stability and reliability were investigated using Analysis of Variance (ANOVA) techniques and by computing the internal consistency reliability for the overall test and the subtests using the Kuder-Richardson Formula 21 (Lord and Novick, 1968).

The data collection and test administration procedures used in this study represent a four-factor completely crossed experimental design. The factors are (1) subjects, the 237 military police trainees who participated in the study; (2) groups, the 2 five round shot groups fired and scored for each table; (3) tables, the 8 distance/position combinations; and (4) repetitions, the 3 repetitions of the 80 shot MPFQC. If these data are treated in a four-factor completely crossed ANOVA and the test is operating as desired, one would expect most of the variance to be accounted for by the subjects. Appreciable variance due to repetitions would indicate a learning (or forgetting) effect. Variance due to tables would indicate non-homogeneous tables. Variance due to groups would indicate a serious lack of stability in the scores.

While an ANOVA appears to be a suitable and straightforward technique for investigating the overall test results, there are several problems which must be addressed. These involve the choice of random and fixed factors, the large number of degrees of freedom involved in testing the statistical significance of several of the F-ratios, and the

similarities and differences in the interpretation of F-ratios and the proportion of variance due to the various main effects and interactions. These problems have been addressed using the MPFQC data as one example by Steinheiser and Epstein (1978).

For this study, subjects, repetitions, and groups were treated as random factors, while tables were treated as a fixed factor. It was necessary to treat repetitions and groups as random factors since it was desirable to consider this particular experiment as a random sample, in time, of the infinite number of times a trainee's competency could be assessed. Treating these factors as fixed would have required any interpretations of the results to be restricted to the rather unrealistic and constrained situation described by this study. On the other hand, the tables were chosen by the Military Police School as its best test of marksmanship. Particularly for a criterion-referenced test where domain specification is so crucial, one must be careful not to over generalize. Therefore, the 8 tables are considered a fixed factor in the ANOVA.

When a large number of degrees of freedom is present in testing F-ratios, it is not difficult to show that main effects and interactions are statistically significant. Since this was the case for this study, the proportion of total variance accounted for by each factor and interaction was computed. The results of the ANOVA were considered both in terms of F-ratios and proportion of variance accounted for by each main effect or interaction.

Because of the mix of fixed and random factors in the ANOVA, it was necessary to compute quasi-F ratios and adjusted values for the

degrees of freedom. Procedures found in Winer (1971) were used to perform the computations. Proportions of total variance accounted for by the main effects and interactions were computed according to procedures published by Dodd and Schultz (1973) and by extending the Cronbach, Gleser, Nanda, and Rajaratnam (1972) procedures for computing variance components used in generalizability studies. The numerical values found using the relatively easy to apply Cronbach, et al. procedures were identical to the Dodd and Schultz results.

#### Measurement Models

Skilled motor performance, such as that described by the MPFQC pistol marksmanship task, should be analyzed by a continuous true score model. Three of the continuous models discussed earlier, the proportion correct, the binomial error, and the beta-binomial Bayesian models, represent a logical progression of increasing complexity and use of information. These models were compared empirically using the MPFQC data. The Rasch model, although a continuous true score model, was not used to analyze these data for several reasons. First, the Rasch model's underlying conception of ability as a latent trait is different from the definition of ability, shared by the other models, as the probability of accomplishing an example of a given task. Thus, defining a meaningful basis of comparison between the two classes of measurement models presents serious problems. Second, the use of the Rasch model, or other latent trait models, presents practical problems, particularly in the area of estimating item parameters. In fact, the number of subjects for whom data were collected may be too small to allow for the estimation of stable item parameters. Finally, how to apply latent

trait theory, in general, and which latent trait model to apply in any given instance, in particular, are matters of considerable current debate. Since it is beyond the scope of this study to enter into that debate, it was decided to limit subsequent empirical analysis to the examples of the class of measurement models based on the properties of the binomial probability distribution, the proportion correct model, the binomial error model, and the beta-binomial Bayesian model.

#### The Proportion Correct Model

The proportion correct model is the simplest of the three, it is also the most "pure" for criterion-referenced testing since no group data are required for its implementation. Pass/fail criteria are based on the statistical properties of the binomial distribution, the criterion for mastery, and the amount of misclassification error that can be tolerated. Given a value for the true ability, the test length, and a criterion passing score, it is possible to compute the probabilities that an individual with the given ability will pass or fail the test. Expected levels of misclassification can be computed for a variety of true abilities, test lengths, and criterion scores. The decision maker must then choose the mix of these factors that best fits the particular testing situation.

Since the MP school uses 70% accuracy for its passing requirement, a true ability of .70 was chosen as the criterion true ability for the analyses conducted in this study. Table 3 shows the probability of misclassification according to the proportion correct model as a function of true ability and criterion score for the 10, 20, 40, and 80 round subtests. For true abilities below .70, the table entry is the probability of a false positive decision. For true abilities at or above

## 10 ROUND SUBTEST

CRITERION SCORE	P(FALSE POSITIVE) GIVEN TRUE ABILITY =				P(FALSE NEGATIVE) GIVEN TRUE ABILITY =				
	.50	.55	.60	.65	.70	.75	.80	.85	.90
ALL PASS	1.0	1.0	1.0	1.0	0	0	0	0	0
1	.999	1.0	1.0	1.0	0	0	0	0	0
2	.989	.995	.998	.999	0	0	0	0	0
3	.945	.973	.988	.995	.002	0	0	0	0
4	.828	.898	.945	.974	.011	.004	.001	0	0
5	.623	.738	.834	.906	.047	.020	.006	.001	0
6	.377	.504	.633	.751	.150	.078	.033	.010	.002
7	.172	.266	.382	.514	.350	.224	.121	.050	.013
8	.055	.100	.167	.262	.617	.474	.322	.180	.070
9	.011	.023	.046	.086	.851	.756	.624	.456	.264
10	.001	.003	.006	.013	.972	.944	.893	.803	.651
ALL FAIL	0	0	0	0	1.0	1.0	1.0	1.0	1.0

Table 3: Proportion Correct Model Probabilities of False Positive and False Negative Misclassification Errors for a Variety of Test Lengths, Criterion Scores, and True Abilities

## 20 ROUND SUBTEST

54

CRITERION SCORE	P(FALSE POSITIVE) GIVEN TRUE ABILITY =				P(FALSE NEGATIVE) NEGATIVE TRUE ABILITY =				
	.50	.55	.60	.65	.70	.75	.80	.85	.90
ALL PASS	1.0	1.0	1.0	1.0	0	0	0	0	0
1									
2									
3	1.0								
4	.999	1.0							
5	.994	.998	1.0						
6	.979	.994	.998	1.0					
7	.942	.979	.994	.998	0				
8	.868	.942	.979	.994	.001	0			
9	.748	.869	.943	.980	.005	.001	0		
10	.588	.751	.872	.947	.017	.004	.001	0	
11	.412	.591	.755	.878	.048	.014	.003	0	
12	.252	.414	.596	.762	.113	.040	.010	.001	0
13	.132	.252	.416	.601	.228	.102	.032	.006	0
14	.058	.130	.250	.417	.392	.214	.087	.022	.002
15	.021	.055	.126	.245	.584	.383	.196	.067	.011
16	.006	.019	.051	.118	.762	.586	.370	.170	.043
17	.001	.005	.015	.044	.893	.775	.589	.352	.133
18	0	.001	.004	.012	.965	.909	.794	.595	.323
19		0	.001	.002	.992	.976	.931	.824	.608
20			0	0	.999	.997	.988	.961	.878
ALL FAIL	0	0	0	0	1.0	1.0	1.0	1.0	1.0

Table 3 (cont)

## 40 ROUND SUBTEST

CRITERION SCORE	P(FALSE POSITIVE) GIVEN TRUE ABILITY =				P(FALSE NEGATIVE) GIVEN TRUE ABILITY =				
	.50	.55	.60	.65	.70	.75	.80	.85	.90
ALL PASS	1.0	1.0	1.0	1.0	0	0	0	0	0
1									
2									
3									
4									
5									
6									
7									
8									
9									
10	1.0								
11	.999	1.0							
12	.997	1.0							
13	.992	.999	1.0						
14	.981	.997	1.0						
15	.960	.991	.999	1.0					
16	.923	.980	.997	1.0					
17	.866	.959	.992	.999	0				
18	.785	.923	.981	.997	0				
19	.682	.867	.961	.992	.001	0			
20	.563	.787	.926	.983	.002	0			

Table 3 (cont)



## 40 ROUND SUBTEST (CONT)

CRITERION SCORE	P(FALSE POSITIVE) GIVEN TRUE ABILITY =				P(FALSE NEGATIVE) GIVEN TRUE ABILITY =				
	.50	.55	.60	.65	.70	.75	.80	.85	.90
21	.437	.684	.870	.964	.006	.001	0	0	0
22	.318	.565	.791	.930	.015	.002	0		
23	.215	.439	.689	.876	.032	.005	0		
24	.134	.319	.568	.798	.063	.012	.001	0	
25	.077	.214	.440	.695	.115	.026	.003	0	
26	.040	.133	.317	.572	.193	.054	.008	0	
27	.019	.075	.211	.441	.297	.103	.019	.001	0
28	.008	.039	.129	.314	.423	.179	.043	.004	0
29	.003	.018	.071	.205	.559	.285	.088	.012	0
30	.001	.007	.035	.121	.691	.416	.161	.030	.001
31	0	.003	.016	.064	.804	.560	.268	.067	.005
32	0	.001	.006	.030	.889	.700	.407	.135	.015
33		0	.002	.012	.945	.818	.563	.244	.042
34		0	.001	.004	.976	.904	.714	.393	.100
35			0	.001	.991	.957	.839	.567	.206
36				0	.997	.984	.924	.737	.371
37				0	.999	.995	.972	.870	.577
38					1.0	.999	.992	.951	.777
39						1.0	.999	.988	.920
40							1.0	.999	.985
ALL FAIL	0	0	0	0	1.0	1.0	1.0	1.0	1.0

Table 3 (cont)

## 80 ROUND SUBTEST

CRITERION SCORE	P(FALSE POSITIVE) GIVEN TRUE ABILITY =				P(FALSE NEGATIVE) GIVEN TRUE ABILITY =				
	.50	.55	.60	.65	.70	.75	.80	.85	.90
ALL PASS	1.0	1.0	1.0	1.0	0	0	0	0	0
.									
.									
25	1.0	1.0	1.0	1.0	0	0	0	0	0
26	.999	1.0							
27	.999	1.0							
28	.998	1.0							
29	.995	1.0							
30	.991	.999	1.0						
31	.984	.999	1.0						
32	.972	.997	1.0						
33	.954	.995	1.0						
34	.937	.991	.999	1.0					
35	.891	.983	.999	1.0					
36	.843	.972	.998	1.0					
37	.783	.954	.995	1.0					
38	.712	.928	.991	1.0					
39	.631	.892	.984	.999	0				
40	.544	.844	.973	.998	0				
41	.456	.785	.956	.996	0				
42	.369	.714	.930	.992	0				
43	.288	.633	.895	.986	.001	0			
44	.217	.546	.848	.975	.002	0			

Table 3 (cont)

## 80 ROUND SUBTEST (CONT)

58

CRITERION SCORE	P(FALSE POSITIVE) GIVEN TRUE ABILITY =				P(FALSE NEGATIVE) GIVEN TRUE ABILITY =				
	.50	.55	.60	.65	.70	.75	.80	.85	.90
45	.157	.457	.789	.959	.003	0	0	0	0
46	.109	.369	.717	.935	.006	0			
47	.073	.288	.636	.900	.012	0			
48	.046	.216	.548	.854	.021	.001	0		
49	.028	.156	.458	.795	.036	.002	0		
50	.016	.108	.369	.724	.059	.005	0		
51	.009	.071	.286	.641	.092	.009	0		
52	.005	.045	.213	.551	.137	.017	.001	0	
53	.002	.027	.152	.458	.195	.029	.001	0	
54	.001	.015	.104	.367	.268	.050	.003	0	
55	.001	.008	.067	.282	.352	.080	.006	0	
56	0	.004	.042	.207	.445	.124	.011	0	
57	0	.002	.025	.145	.542	.182	.022	0	
58	0	.001	.014	.097	.637	.255	.039	.001	0
59		0	.007	.061	.725	.343	.066	.003	0
60		0	.004	.037	.802	.440	.107	.006	0
61		0	.002	.021	.865	.543	.163	.013	0
62		0	.001	.011	.913	.644	.238	.026	0
63			0	.006	.947	.736	.329	.048	.001
64			0	.003	.970	.816	.434	.084	.002
65			0	.001	.984	.879	.545	.138	.005
66				0	.992	.926	.654	.213	.012

Table 3 (cont)

## 80 ROUND SUBTEST (CONT)

CRITERION SCORE	P(FALSE POSITIVE) GIVEN TRUE ABILITY =				P(FALSE NEGATIVE) GIVEN TRUE ABILITY =				
	.50	.55	.60	.65	.70	.75	.80	.85	.90
67	0	0	0	0	.994	.958	.753	.309	.027
68				0	.998	.978	.836	.424	.054
69				0	.999	.989	.899	.548	.100
70					1.0	.995	.944	.670	.173
71					1.0	.998	.971	.779	.277
72					1.0	.999	.987	.866	.407
73						1.0	.995	.927	.554
74						1.0	.998	.965	.700
75						1.0	.999	.986	.823
76							1.0	.995	.912
77							1.0	.999	.965
78								1.0	.989
79								1.0	.998
80									1.0
ALL FAIL	0	0	0	0	1.0	1.0	1.0	1.0	1.0

Table 3 (cont)

.70, the table entry is the probability of a false negative decision.

In interpreting and using Table 3, it is important to remember that each misclassification probability refers only to the true ability represented by the entries in any particular column of the table. For example, for test length equals 10 items, true ability equals .70, and criterion score equals 7, the false negative probability given in the table is .350. This means that if all examinees had true abilities of .70, approximately one-third of them would be expected to fail a ten item test with a passing criterion score of seven correct. In most cases the examinee group will not consist of individuals all of whom have the same true ability. Therefore, in using Table 3, the decision maker must consider a mix of abilities, and the expected false positive and false negative misclassifications associated with that mix, in choosing a criterion score.

In choosing criterion scores for subsequent analysis, the absolute and relative misclassification probabilities for the range of true abilities, .50 to .90, represented in Table 3 were simultaneously considered. The relative values of the false positive and false negative error probabilities are important since the losses associated with each type of error are being treated as equal. The scores chosen should yield the lowest absolute error probabilities and the closest relative error probabilities across the range of ability levels considered.

#### The Binomial Error Model

The binomial error model builds on the basic foundation of the proportion correct model by incorporating observed group data into the decision making process. There is less subjective judgment in weighing

alternatives required by this approach than is the case for the proportion correct model, however, the statistical model underlying it must be appropriate. The binomial error model was implemented by completing the following steps. First, the observed scores for each subtest were analyzed to determine whether the distributions were statistically significantly different from negative hypergeometric distributions. This required computing the mean, variance, and Kuder-Richardson Formula 21 reliability for each subtest, and solving for the parameters of its associated negative hypergeometric distribution using equations (19), (20), and (21). A chi-square goodness of fit test was applied to determine whether or not the model was appropriate for the data. Estimated true scores corresponding to each observed score were then computed using the regression equation shown in equation (22). The criterion observed score was the lowest score that yielded an estimated true score greater than or equal to .70. Table 4 shows the chi-square probabilities that the observed scores represent samples of scores from negative hypergeometric distributions, the recommended criterion scores, and the associated estimated true scores for each of the subtests.

To compute the expected misclassification under the binomial error model the following procedure was employed. Since the overall true score distribution is a member of the beta family, it was assumed that the error of estimation around each estimated true score was also a member of the beta family with a mean equal to the estimated true score and variance equal to  $\sigma_B^2 (1 - \alpha_{21})$ ; where  $\sigma_B^2$  is the estimated variance of the true score distribution. The above equation is the analogue of the usual equation for the variance of the error of estimate for classical

SUBTEST	$P(\chi^2)$	CRITERION SCORE	ESTIMATED TRUE SCORE	SUBTEST	$P(\chi^2)$	CRITERION SCORE	ESTIMATED TRUE SCORE
TABLE11	>.25	8	.756	TABLE15	>.03	6	.704
TABLE12	>.25	8	.731	TABLE16	>.05	7	.747
TABLE13	>.75	8	.716	TABLE17	>.90	6	.717
TABLE14	>.05	8	.748	TABLE18	>.10	5	.728
TABLE21	>.75	8	.757	TABLE25	>.75	7	.753
TABLE22	>.50	8	.749	TABLE26	>.50	7	.733
TABLE23	>.05	8	.715	TABLE27	>.05	7	.753
TABLE24	>.50	8	.762	TABLE28	>.25	4	.723
TABLE31	>.75	8	.754	TABLE35	>.50	6	.724
TABLE32	>.50	8	.762	TABLE36	>.98	7	.756
TABLE33	>.50	8	.742	TABLE37	>.25	6	.719
TABLE34	>.25	7	.705	TABLE38	>.75	4	.718
HARD21	>.75	15	.717	EASY21	>.10	13	.729
HARD22	<.01	14	.720	EASY22	>.75	12	.714
HARD23	<.01	14	.719	EASY23	>.75	12	.704
HARD24	<.01	14	.721	EASY24	>.95	13	.724
HARD25	>.10	15	.731	EASY25	>.98	12	.720
HARD26	>.25	15	.725	EASY26	>.75	11	.714

Table 4: Binomial Error Model  $\chi^2$  Probabilities that Subtest Scores Represent Samples from a Negative Hypergeometric Distribution, and Criterion Observed and Estimated True Scores

SUBTEST	P( $\chi^2$ )	CRITERION SCORE	ESTIMATED TRUE SCORE	SUBTEST	P( $\chi^2$ )	CRITERION SCORE	ESTIMATED TRUE SCORE
MIX201	>.95	14	.723	MIX207	>.98	14	.720
MIX202	>.95	14	.728	MIX208	>.10	14	.716
MIX203	>.25	14	.719	MIX209	>.75	14	.729
MIX204	>.50	14	.719	MIX210	>.03	13	.710
MIX205	>.50	13	.700	MIX211	>.25	13	.711
MIX206	>.99	14	.720	MIX212	>.25	14	.730
HARD41	>.75	29	.708	EASY41	>.10	26	.710
HARD42	>.50	29	.711	EASY42	>.10	27	.718
HARD43	>.10	29	.716	EASY43	>.99	25	.701
MIX41	>.90	28	.717	MIX44	>.50	28	.711
MIX42	>.98	28	.712	MIX45	>.50	27	.703
MIX43	>.25	28	.715	MIX46	>.25	27	.706
REP1	>.95	56	.709	REP3	>.25	55	.703
REP2	>.25	56	.708				

Table 4 (cont)



regression. Given the mean and variance of a beta distribution one can compute its parameter values,  $a$  and  $b$  by solving the following equations simultaneously:  $\mu_B = a/(a+b)$ ,  $\sigma_B^2 = ab/[(a+b+1)^2(a+b+2)]$  (Novick and Jackson, 1974, p.113). This distribution describes the error of estimation. The expected false negative rate for each failing score is the area of the distribution above ability = .70. That is, the probability that ability  $\geq .70$  even though a decision to fail is implied by the estimated true score. Similarly, the false positive rate for each passing score is found by computing the area of the distribution below ability = .70.

#### The Beta-Binomial Bayesian Model

The beta-binomial Bayesian model uses prior information about the overall abilities of the examinees (or similar examinees) as its starting point for computing recommended criterion passing scores. The model assumes that examinee abilities are distributed as a beta distribution. Once that distribution is identified, by determining its parameter values, then each examinee's observed score can be interpreted as an indicator of the beta distribution which best describes the ability group to which that individual belongs. This is simply because given a prior ability distribution which is a member of the beta family,  $B(a,b)$ , and an observed score,  $x$  correct of  $n$  items, which is part of a binomial distribution, the posterior ability distribution is also a member of the beta family,  $B(a+x, b+n-x)$ . The mean of the posterior ability distribution also provides a true ability estimate.

In order to use the model to find a criterion passing score, the areas above and below the criterion true ability, in this case .70,

corresponding to the posterior ability distributions for each observed score are computed. The observed score that provides a criterion passing score with equal false positive and false negative losses associated with it is the lowest score for which the area of the posterior distribution above the criterion ability is .50 (Novick and Lewis, 1974). These areas can also be interpreted as misclassification probabilities.

If the observed score is below the criterion passing score, the area under the curve above .70 is the probability of a false negative. This is because, the area above .70 is the probability that the individual's true ability is .70 or better, computed on a curve associated with a fail decision. Conversely, if the observed score is at or above the criterion passing score, the area under the curve below .70 is the probability of a false positive decision. The results of the analyses described above for these data are found in Table 5.

The prior ability distribution used for the analyses in this study was obtained by asking MP school marksmanship instructors to estimate the distribution of scores that would be obtained on the test by a hypothetical group of thirty students. The group of eleven "experts" were asked to fill out a form for each of the eight tables on the MPFQC requesting them to estimate how many of the thirty hypothetical students would get 0-1, 2-3, 4-5, 6-7, 8-9, or 10 hits. In the analysis, these data were combined to yield an average prior estimate of an observed score distribution on the MPFQC. Using results obtained by Lord and Novick (1968, p.522), it is possible to relate the moments of the observed score distribution to the moments of the underlying true beta ability distribution:

## 10 ROUND SUBTESTS

SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )
0	.002	4	.042	8	.713
1	.002	5	.118	9	.888
2	.004	6	.266	10	.975
3	.013	7	.482		

## 20 ROUND SUBTESTS

SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )
0-6	.002	11	.091	16	.815
7	.003	12	.182	17	.917
8	.006	13	.317	18	.972
9	.016	14	.487	19	.993
10	.041	15	.664	20	.999

Table 5: Beta-Binomial Bayesian Model Probabilities that Ability  $\geq .70$   
as a Function of Observed Score and Prior Distribution:  
Prior Based on All MPFQC Tables

## 40 ROUND SUBTESTS

SCORE	PROBABILITY (ABILITY $\geq$ .7)	SCORE	PROBABILITY (ABILITY $\geq$ .7)	SCORE	PROBABILITY (ABILITY $\geq$ .7)
0-17	.002	25	.160	32	.911
18	.003	26	.250	33	.956
19	.004	27	.363	34	.981
20	.007	28	.491	35	.993
21	.014	29	.622	36	.998
22	.028	30	.742	37	.999
23	.053	31	.840	38-40	>.999
24	.096				

## 80 ROUND SUBTESTS

SCORE	PROBABILITY (ABILITY $\geq$ .7)	SCORE	PROBABILITY (ABILITY $\geq$ .7)	SCORE	PROBABILITY (ABILITY $\geq$ .7)
0-41	.002	52	.170	61	.885
42-43	.003	53	.235	62	.927
44	.005	54	.313	63	.956
45	.007	55	.400	64	.975
46	.012	56	.494	65	.987
47	.019	57	.588	66	.994
48	.032	58	.679	67	.997
49	.051	59	.761	68	.999
50	.079	60	.830	69-80	>.999
51	.118				

Table 5 (cont)

## 10 ROUND SUBTESTS

SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )
0-1	<.001	5	.069	8	.511
2	.001	6	.161	9	.714
3	.007	7	.312	10	.871
4	.024				

## 20 ROUND SUBTESTS

SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )
0-6	<.001	11	.056	16	.672
7	.001	12	.116	17	.811
8	.003	13	.213	18	.909
9	.009	14	.448	19	.965
10	.024	15	.509	20	.989

## 40 ROUND SUBTESTS

SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )
0-18	<.001	26	.179	34	.953
19	.001	27	.271	35	.979
20	.003	28	.483	36	.992
21	.007	29	.507	37	.997
22	.016	30	.632	38	.999
23	.034	31	.746	39	>.999
24	.063	32	.839	40	>.999
25	.110	33	.908		

Table 5 (cont): Prior Based on Hard MPFQC Tables

## 10 ROUND SUBTESTS

SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )
0-1	.001	5	.148	8	.788
2	.003	6	.423	9	.935
3	.015	7	.560	10	.991
4	.053				

## 20 ROUND SUBTESTS

SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )
0	<.001	11	.110	16	.859
1-6	.001	12	.215	17	.944
7	.002	13	.365	18	.984
8	.006	14	.545	19	.997
9	.019	15	.721	20	>.999
10	.049				

## 40 ROUND SUBTESTS

SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )	SCORE	PROBABILITY (ABILITY $\geq .7$ )
0	<.001	24	.110	31	.868
1-18	.001	25	.183	32	.930
19	.003	26	.280	33	.968
20	.007	27	.401	34	.987
21	.015	28	.533	35	.996
22	.032	29	.663	36	.999
23	.061	30	.778	37-40	>.999

Table 5 (cont): Prior Based on Easy MPFQC Tables

$$\mu_B = \mu_x/n, \quad (29)$$

$$\sigma_B^2 = 1/[n(n-1)] [\sigma_x^2 - (1/n)\mu_x(n-\mu_x)] \quad (30)$$

where  $\mu_B$  is the mean of the beta distribution,  $\sigma_B^2$  is the variance of the beta distribution,  $\mu_x$  is the mean of the observed score distribution,  $\sigma_x^2$  is the variance of the observed score distribution, and  $n$  is the number of test items. These same equations were used to compute the moments, and subsequently the parameters, of the prior estimated beta distribution. The estimated observed score distribution is described in Table 6. The resulting prior beta distribution was  $B(2.797, 1.498)$  for the whole test. The prior distribution for the four hard tables was  $B(4.654, 3.534)$ . The prior distribution for the four easy tables was  $B(2.424, 0.878)$ .

#### Comparing the Models

The models were compared on the basis of the accuracy of the master/nonmaster classifications that followed from each model's recommended criterion passing score and the accuracy of each model's estimated true scores. Two summary statistics were computed for each subtest to assess the accuracy of the estimated true scores. An absolute discrepancy index was defined and computed as

$$\sum_{i=1}^k (ET_{ij} - T_i), \quad (31)$$

where  $ET_{ij}$  is the estimated true score for person  $i$  on subtest  $j$ ,  $T_i$  is the criterion true score for person  $i$ , and  $k$  is the number of persons.

A squared discrepancy index was defined and computed as

$$\sum_{i=1}^k (ET_{ij} - T_i)^2. \quad (32)$$

SCORE	AVERAGE PREDICTED FREQUENCIES							
	TABLE1	TABLE2	TABLE3	TABLE4	TABLE5	TABLE6	TABLE7	TABLE8
0-1	1.31	1.09	1.45	1.18	1.18	2.36	.91	.36
2-3	4.18	2.55	3.91	2.42	1.27	2.36	.82	.55
4-5	8.89	8.45	10.00	8.13	5.73	5.00	1.69	.73
6-7	10.20	10.45	9.73	11.04	10.91	8.46	9.04	2.18
8-9	4.52	6.64	3.73	6.45	9.36	10.00	10.39	7.91
10	.89	.82	1.18	.78	1.55	1.82	7.15	18.27

AVERAGE PREDICTED SCORE (ALL TABLES): 6.51

PREDICTED VARIANCE (ALL TABLES): 6.13

PRIOR BETA DISTRIBUTION (ALL TABLES): B(2.797,1.498)

AVERAGE PREDICTED SCORE (HARD TABLES): 5.68

PREDICTED VARIANCE (HARD TABLES): 4.86

PRIOR BETA DISTRIBUTION (HARD TABLES): B(4.654,3.534)

AVERAGE PREDICTED SCORE (EASY TABLES): 7.34

PREDICTED VARIANCE (EASY TABLES): 6.03

PRIOR BETA DISTRIBUTION (EASY TABLES): B(2.424,0.878)

**Table 6: Expected Examinee Performance on the Military Police  
Firearms Qualification Course and Implied Prior Beta  
Distributions  
(Data Represents Opinions of 11 Military Police School  
Instructors)**



In order to determine the classification accuracy, a series of 2x2 tables showing the relationship between the criterion master/nonmaster decisions based on the 240 round total test or the 120 round hard and easy tests and the subtest master/nonmaster decisions based on each recommended criterion score were constructed. Figure 2 describes the properties of the cells and the marginals of such a table. The probability of a false positive equals  $P(\text{pass subtest and nonmaster})$ , the probability of a false negative equals  $P(\text{fail subtest and master})$ .

Tables such as that in Figure 2 summarize more detailed tables which relate each score on a subtest to the true master/nonmaster classification. The probability of passing a subtest is actually the sum of the probabilities of passing the subtest for each score above the criterion. A similar relationship exists for failing scores. These relationships can be written

$$P(\text{pass subtest}) = \sum_{x=c}^n P(\text{obtain score } x) = \quad (33)$$

$$\sum_{x=c}^n P(\text{obtain score } x \text{ and master}) +$$

$$\sum_{x=c}^n P(\text{obtain score } x \text{ and nonmaster}),$$

$$P(\text{fail subtest}) = \sum_{x=0}^{c-1} P(\text{obtain score } x) = \quad (34)$$

$$\sum_{x=0}^{c-1} P(\text{obtain score } x \text{ and master}) +$$

$$\sum_{x=0}^{c-1} P(\text{obtain score } x \text{ and nonmaster}),$$

where,  $c$  is the criterion passing score and  $n$  is the number of items on the subtest.

"TRUE" CLASSIFICATION				
		MASTER	NON MASTER	
SUBTEST DECISION	PASS	PASS SUBTEST and MASTER	PASS SUBTEST and NON MASTER	PASS SUBTEST
	FAIL	FAIL SUBTEST and MASTER	FAIL SUBTEST and NON MASTER	FAIL SUBTEST
		MASTER	NON MASTER	

$$P(\text{PASS SUBTEST}) = P(\text{PASS SUBTEST and MASTER}) + P(\text{PASS SUBTEST and NON MASTER})$$

$$P(\text{FAIL SUBTEST}) = P(\text{FAIL SUBTEST and MASTER}) + P(\text{FAIL SUBTEST and NON MASTER})$$

Figure 2: True Classification versus Subtest Classification  
Contingency Matrix

Since it is of interest to compare the models with respect to the difference in misclassification observed and the misclassification predicted by the model, as well as on the basis of absolute observed misclassification, some common index must be found. The previous discussions of the characteristics of the models have shown that the model's expected misclassification probabilities are conditional probabilities. For the proportion correct model the expected misclassification is obtained by summing the appropriate terms of  $P(\text{obtain score } x \text{ given ability})$ . For the binomial error and Bayesian models the misclassification is computed as a function of  $P(\text{ability given score } x)$ . These conditional probabilities and the probabilities shown in Figure 2 are related through the definition of conditional probability

$$P(A|B) = P(A \text{ and } B)/P(B), \text{ and} \quad (35)$$

$$P(B|A) = P(A \text{ and } B)/P(A), \quad (36)$$

or in terms of a testing situation

$$P(\text{master} | \text{score} = x) = P(\text{master and score} = x)/P(\text{score} = x), \quad (37)$$

$$P(\text{score} = x | \text{master}) = P(\text{master and score} = x)/P(\text{master}), \quad (38)$$

$$P(\text{nonmaster} | \text{score} = x) = P(\text{nonmaster and score} = x)/P(\text{score} = x) \quad (39)$$

and

$$P(\text{score} = x | \text{nonmaster}) = P(\text{nonmaster and score} = x)/P(\text{nonmaster}). \quad (40)$$

The discussion above suggests the following scheme for comparing the model's classification accuracies. First, empirically determine the observed false positive and false negative probabilities by dividing the number of subjects in the (PASS SUBTEST and NONMASTER) and the (FAIL SUBTEST and MASTER) cells of Figure 2 by the sample size (=237). Since these numbers will vary with the criterion passing score, any differ-

ences in the model's recommended passing scores will be reflected. For the proportion correct model

$$P(\text{obtain score } x \geq \text{criterion} | \text{ability} = \text{nonmaster}) = \quad (41)$$

$$\frac{P(\text{obtain score } x \geq \text{criterion and ability} = \text{nonmaster})}{P(\text{ability} = \text{nonmaster})}, \text{ and}$$

$$P(\text{obtain score } x < \text{criterion} | \text{ability} = \text{master}) = \quad (42)$$

$$\frac{P(\text{obtain score } x < \text{criterion and ability} = \text{master})}{P(\text{ability} = \text{master})}$$

are the expected false positive and false negative rates. Therefore, each conditional probability on the left of the equations above must be multiplied by the probability of the ability in the sample. When these terms are summed for all passing and failing scores, indices comparable to  $P(\text{pass subtest and nonmaster})$  and  $P(\text{fail subtest and master})$  are obtained.

For the binomial error and Bayesian models

$$P(\text{ability} = \text{nonmaster} | \text{score} \geq \text{criterion}) = \quad (43)$$

$$\frac{P(\text{ability} = \text{nonmaster and score} \geq \text{criterion})}{P(\text{score} \geq \text{criterion})}, \text{ and}$$

$$P(\text{ability} = \text{master} | \text{score} < \text{criterion}) = \quad (44)$$

$$\frac{P(\text{ability} = \text{master and score} < \text{criterion})}{P(\text{score} < \text{criterion})}$$

are the expected false positive and false negative rates. Therefore, each conditional probability on the left of the equations above must be multiplied by the probability of obtaining the score in the sample. When these terms are summed for all passing and failing scores, indices comparable to  $P(\text{pass subtest and nonmaster})$  and  $P(\text{fail subtest and master})$  are obtained.

#### 4. RESULTS

The description of the results of this study is divided into three sections. The first section addresses the MPFQC performance data. Descriptive statistics for the 240 round and 120 round criterion tests and the sampled subtests are reported and the results of the ANOVA are described. The implications of these results for interpreting the characteristics of the MPFQC as a testing instrument are discussed.

The second and third sections address the comparisons of the models. The second section refers to the results based on the 240 round criterion test, and the third section refers to the results based on the 120 round hard and easy criterion tests. The models are compared with respect to their recommended criterion scores, the amount of misclassification observed, and the accuracy with which the models estimated the amount of misclassification. The results relating to the accuracy of the models in estimating examinee true scores are then presented.

##### Characteristics of the MPFQC Performance Data

Results for the total test of 240 rounds indicate that it is a reliable, moderately difficult test of marksmanship. The scores form a negatively skewed single peak distribution (Figure 3). The mean score is 184.591 (76.9%), the median is 185.800 (77.4%), and the mode is 185 (77.1%). The test scores have a variance of 614.336 and a KR-21 reliability of .934.

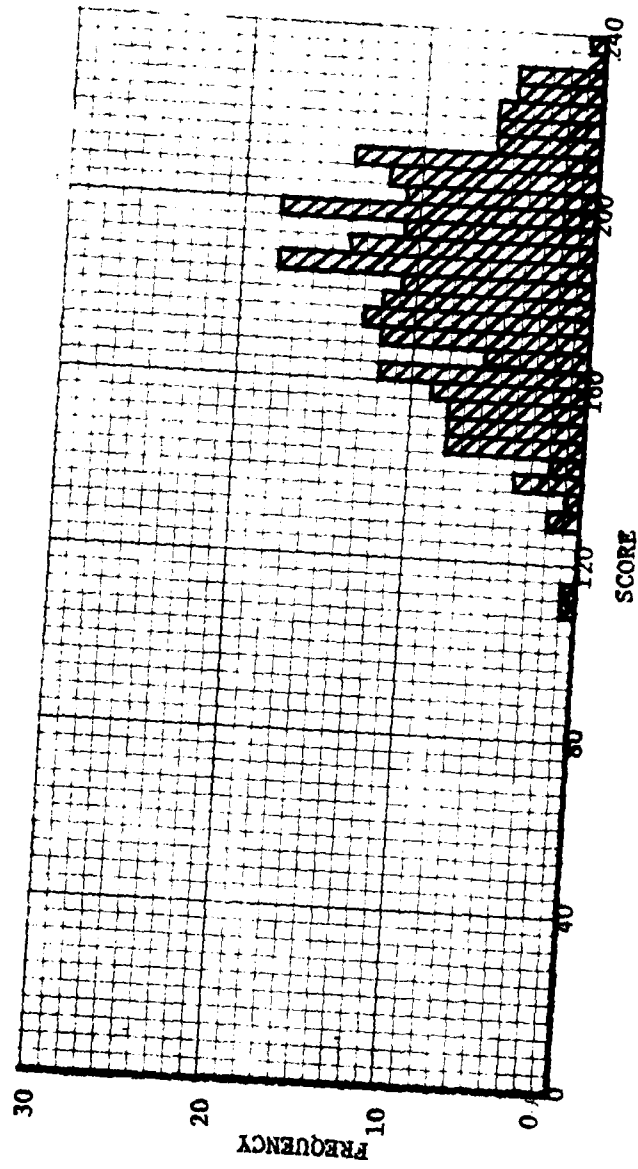


Figure 3: Distribution of Scores on 240 Round Criterion Test

The three 80 round tests (Rep1, Rep2, Rep3) show similar characteristics (Figure 4). Rep1 has a mean score of 60.447 (75.6%), a median of 60.321 (75.4%), and a mode of 60 (75%). Its variance is 89.952 and its KR-21 reliability is .846. Rep2 has a mean score of 60.861 (76.1%), median of 62.107 (77.6%), and mode of 62 (77.5%). Its variance is 107.824 and it has a KR-21 reliability of .875. Rep3 has a mean score of 63.283 (79.1%), median of 63.906 (79.9%), and a mode of 64 (80%). Its variance is 85.271 and its KR-21 reliability is .855. These data suggest that the test became slightly easier with each repetition, perhaps reflecting a practice effect. The change in difficulty was further explored in an analysis of variance, described below, which showed that the effect was not sufficiently large to disturb the interpretation of the results.

If one breaks the data into the scores representing the sum of the 10 round hard subtests (Tables 11 - 14, 21 - 24, 31 - 34) and the sum of the 10 round easy subtests (Tables 15 - 18, 25 - 28, 35 - 38), the fact that the MPFQC is actually made up of two rather different tests becomes obvious. The hard test of 120 rounds (Figure 5) has a mean of 77.253 (64.4%), median of 76.417 (63.7%), and mode of 61 (50.8%). Its distribution has a slight positive skew. Its variance and KR-21 reliability are 299.630 and .915 respectively. The 120 round easy test (Figure 6) has a mean of 107.338 (89.4%), median of 109.909 (91.6%), and mode of 114 (95%). Its distribution is negatively skewed, it has a variance of 88.148, and its KR-21 reliability is .878. Further, cross-tabulations of the pass/fail decisions based on the 120 round tests and the total 240 round test show that no one who failed the easy test pass-

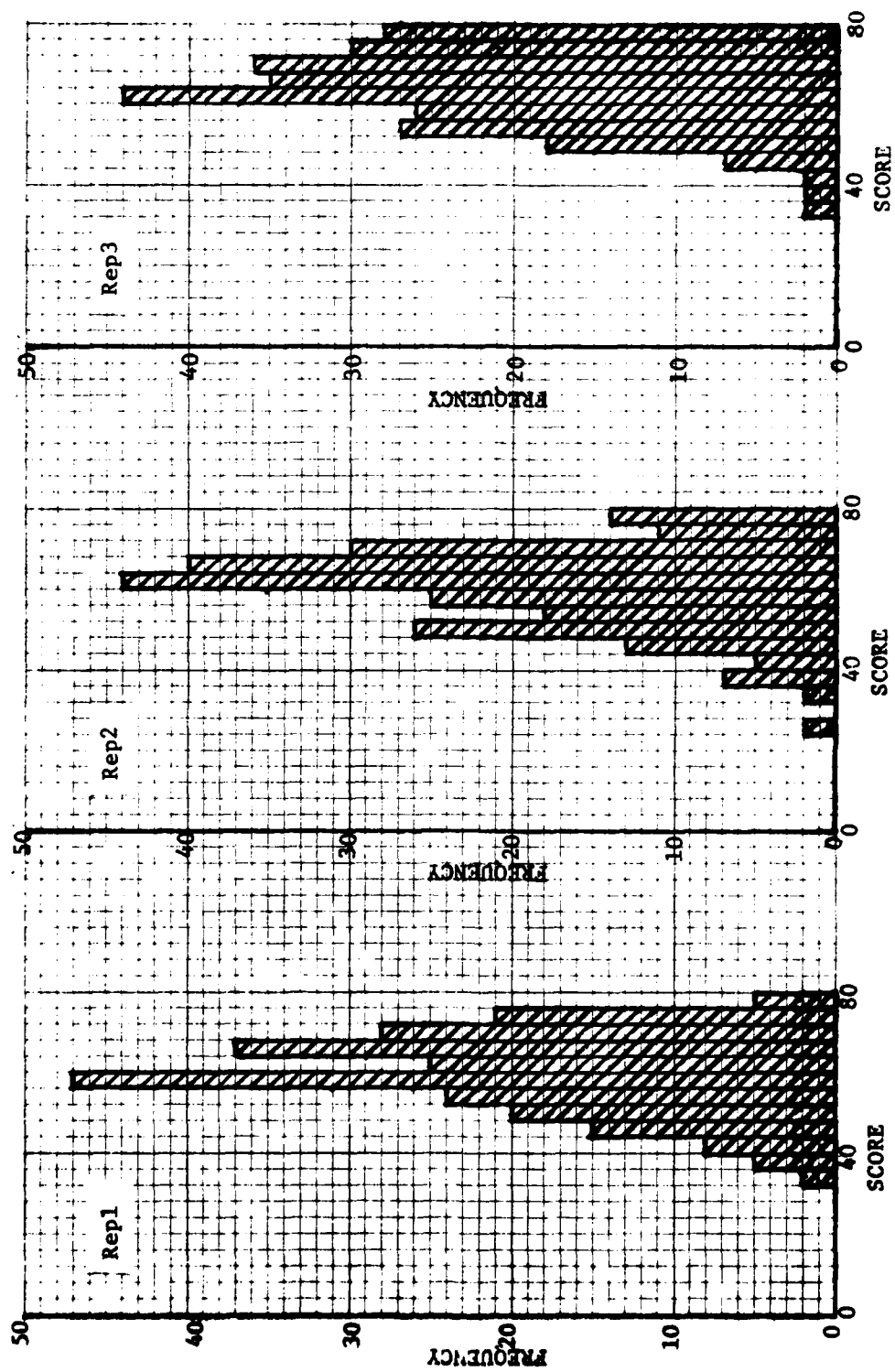


Figure 4: Distribution of Scores on 80 Round Subtests



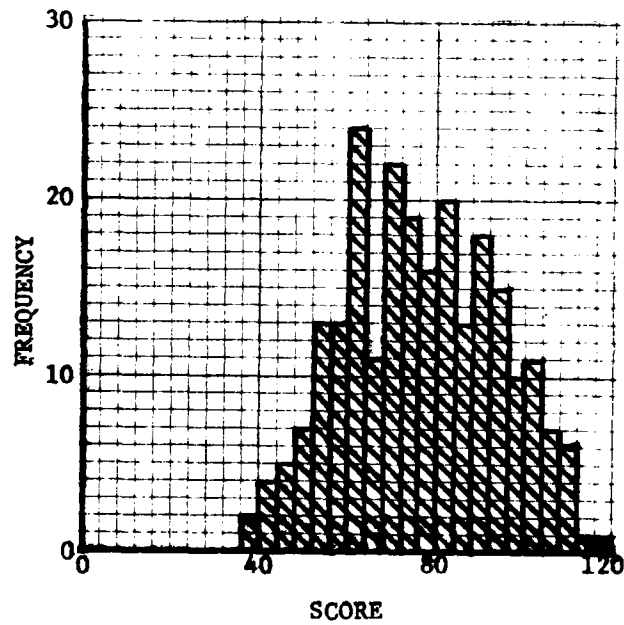


Figure 5: Distribution of Scores on 120 Round Hard Criterion Test

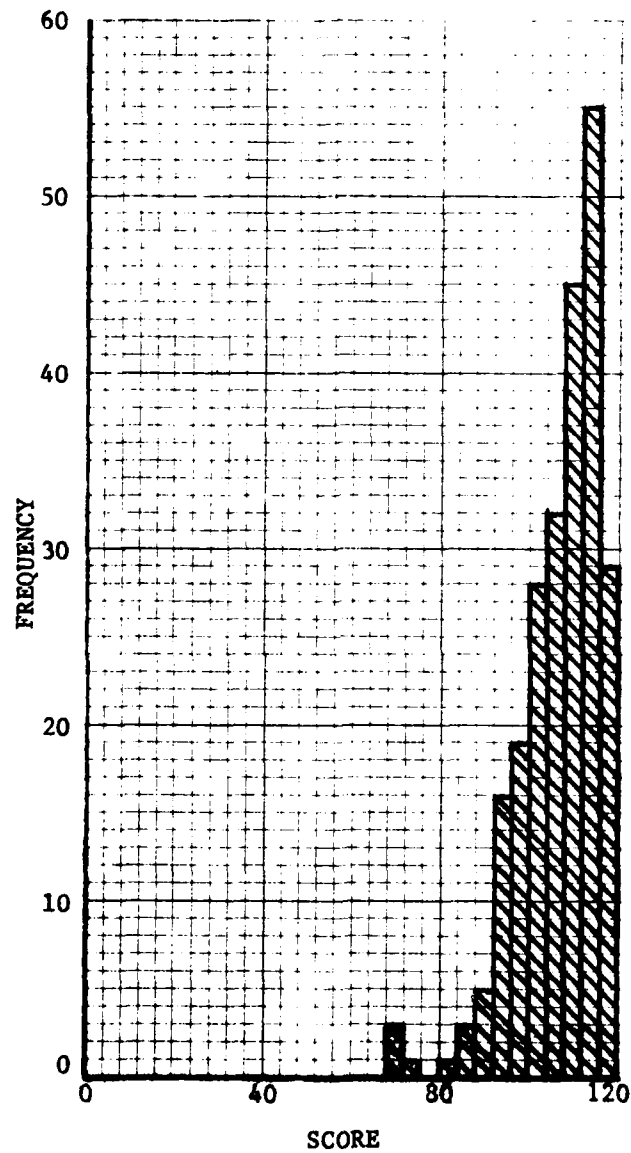


Figure 6: Distribution of Scores on 120 Round Easy Criterion Test

ed the hard test, no one who passed the hard test failed the easy test, no one who failed the 240 round test passed the 120 round hard test, and no one who failed the 240 round test failed the easy test. These data suggest that the MPFQC is not measuring a unitary skill. Rather, there appear to be two distinct skills being demonstrated. One is a general ability to shoot accurately, regardless of the distance (from 7 to 35 meters) from the target. A second skill is demonstrated as an ability to shoot accurately only at short distances (7 and 15 meters), without that skill consistently being shown at the longer distances from the target.

Table 2 summarizes the descriptive data for the 10, 20, and 40 round subtests. These data corroborate the results for the longer tests. The subtests made up of the difficult tables are consistently more difficult than those which represent a mix of the hard and easy tables which are, in turn, more difficult than the tests made up exclusively of the easy tables. The mean scores for the 10 round hard subtests vary from 5.37 (53.7%) to 7.15 (71.5%), and the means for the easy subtests vary from 8.23 (82.3%) to 9.76 (97.6%). For the 20 round hard subtests the means vary from 12.26 (61.3%) to 13.54 (67.7%), for the easy subtests they vary from 17.66 (88.3%) to 18.28 (91.4%), and they vary from 14.68 (73.4%) to 15.88 (79.4%) for the mix subtests. For the 40 round subtests the means are 25.08 (62.7%) to 26.8 (67%) for the hard subtests, 35.36 (88.4%) to 36.52 (91.3%) for the easy subtests, and 29.96 (74.9%) to 30.76 (76.9%) for the mix subtests. These data are very consistent within a type of subtest and show clear differences

between subtest types, regardless of the number of rounds included. None of the ranges of means for different subtest types overlap. The KR-21 reliabilities are consistently acceptable considering the relatively short test lengths involved. The lowest reliabilities are .438 and .448 for two of the very easy 10 round subtests. Reliabilities as high as .794 are found for the 10 round subtests. Twenty round subtest reliabilities vary from .551 to .778, and 40 round subtest reliabilities vary from .727 to .843. With the exception of the low reliabilities for some of the easy 10 round subtests, no particular patterns for the reliability data are evident.

Figure 7 shows these data plotted as test characteristic curves. The curves show the cumulative proportions of the examinees achieving each score. The curves further illustrate the characteristics of the MPFQC discussed thus far. That is, the curves are remarkably similar to one another within a test type and clearly distinct between test types. This is found regardless of test length. It is also clear from the curves that the 80 round subtests and 240 round criterion test are similar to the other mix subtests. The data also show that the 120 round hard and easy tests are similar to the other subtests of their respective types.

In general, the MPFQC appears to be a reliable, easily interpretable measure of pistol marksmanship. When data from all eight tables are considered in a composite test, the scores provide a general index of marksmanship. A more fine grained index of marksmanship ability is available by considering scores from the hard and easy tables independently.

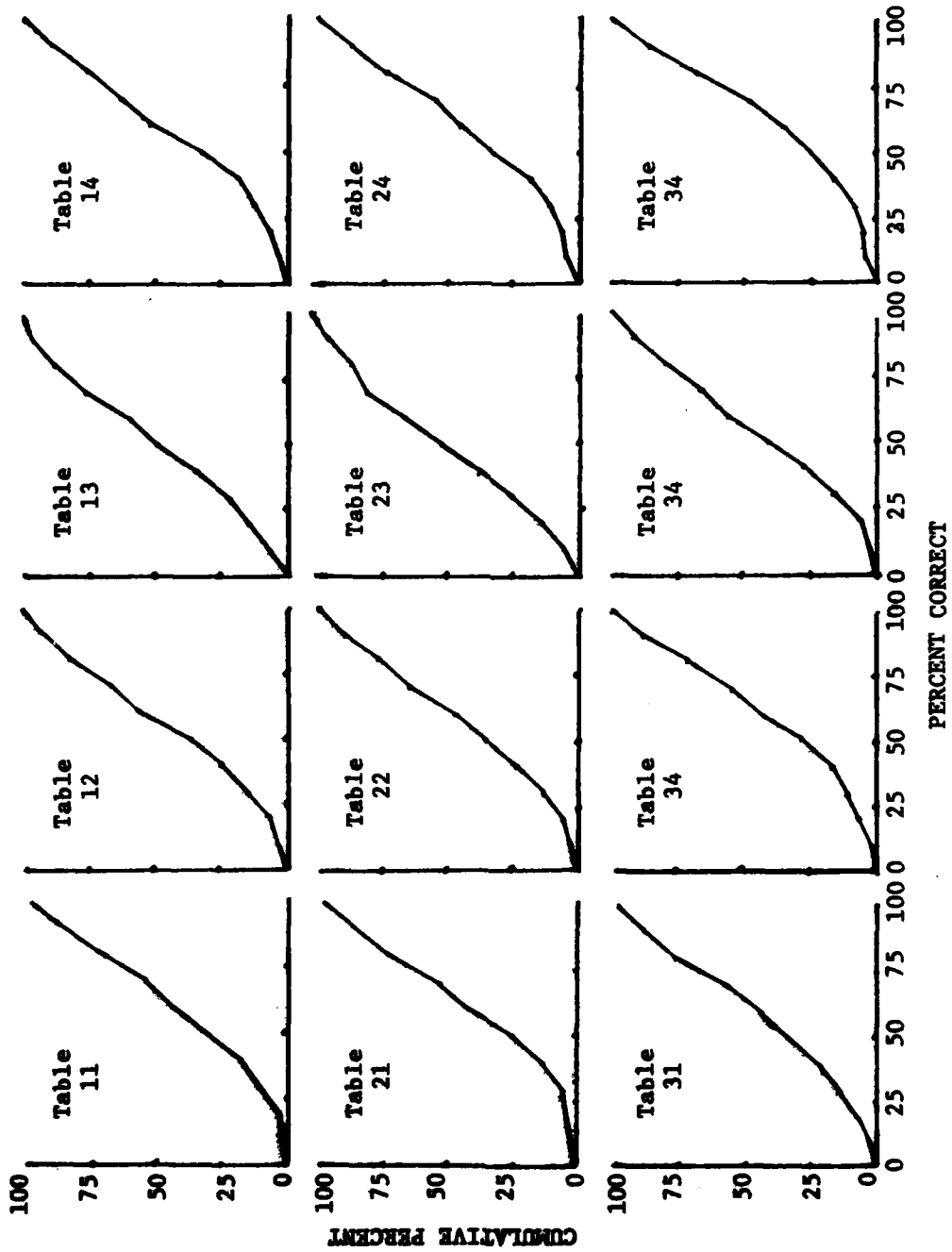


Figure 7: Test Characteristic Curves: 10 Round Hard Subtests

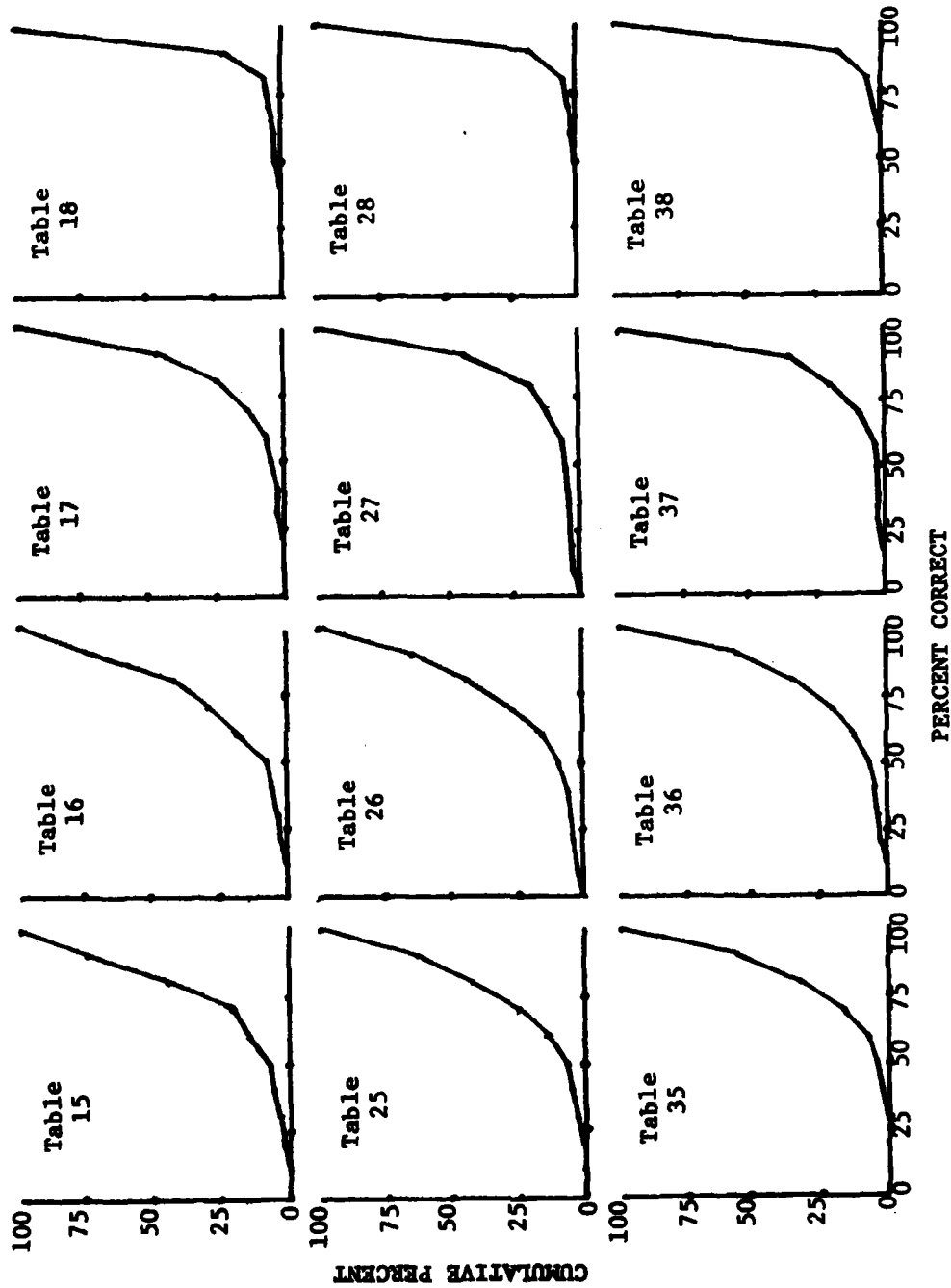


Figure 7 (cont): 10 Round Easy Subtests

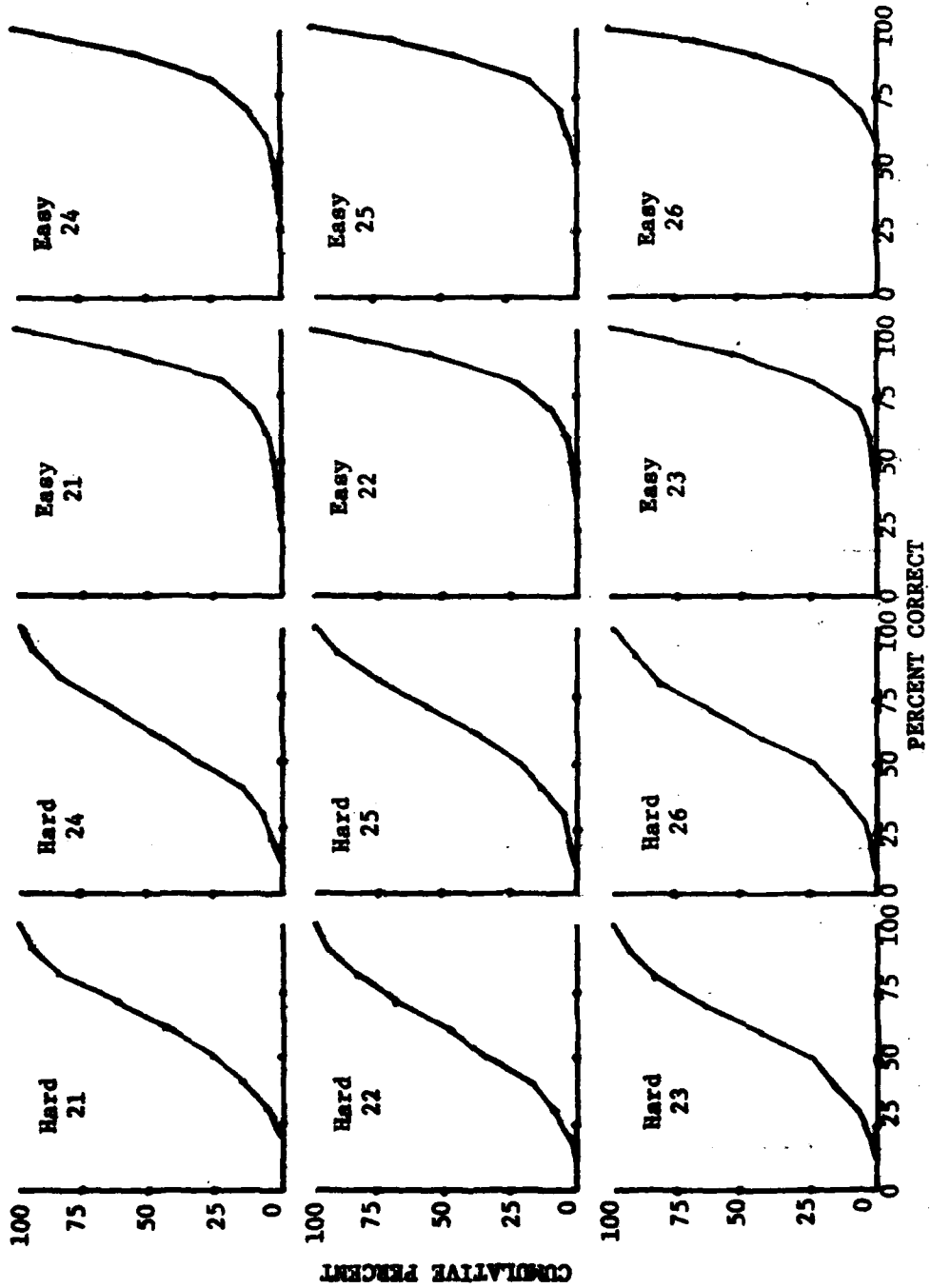


Figure 7 (cont): 20 Round Hard and Easy Subtests

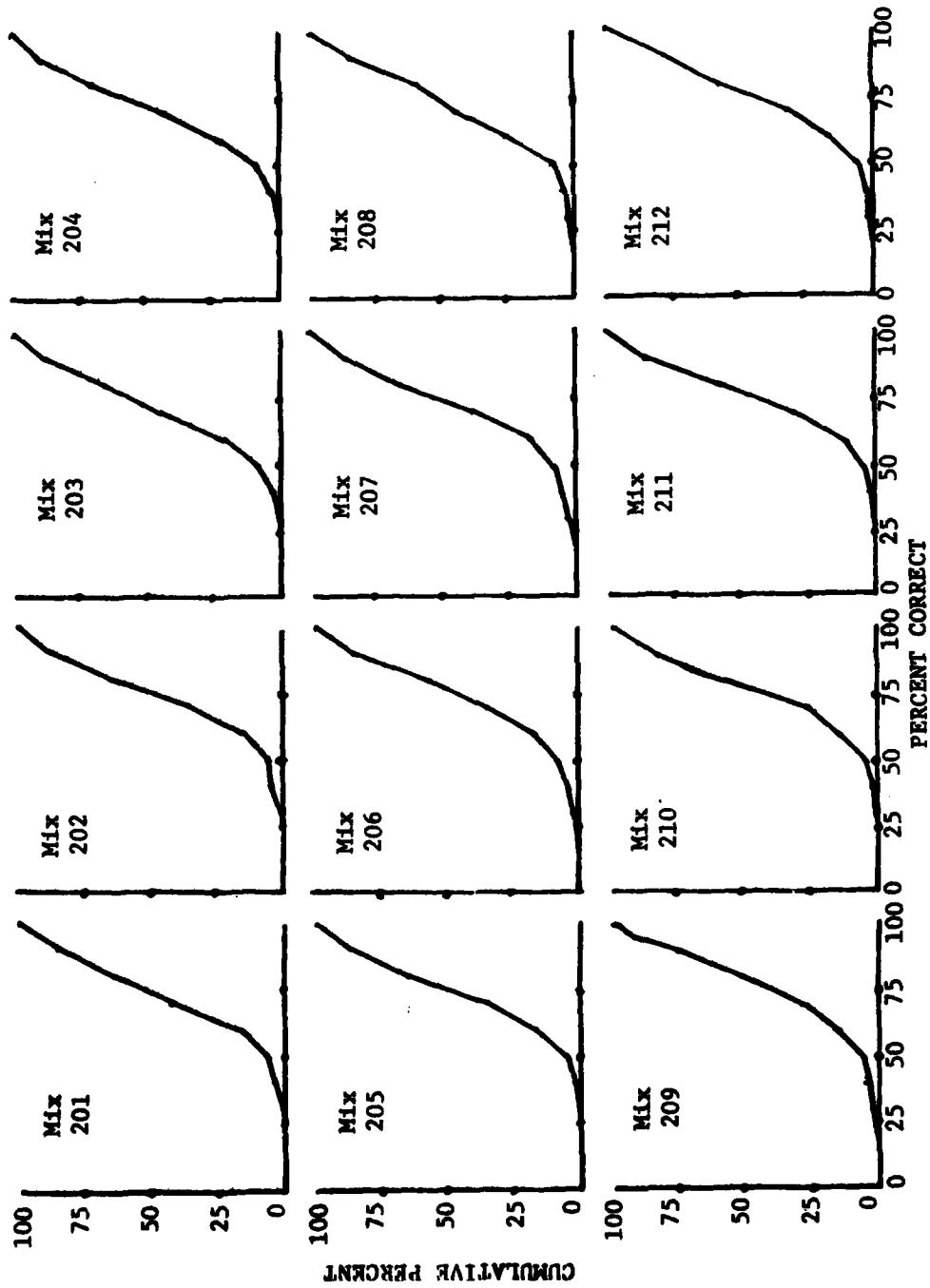


Figure 7 (cont): 20 Round Mix Subtests



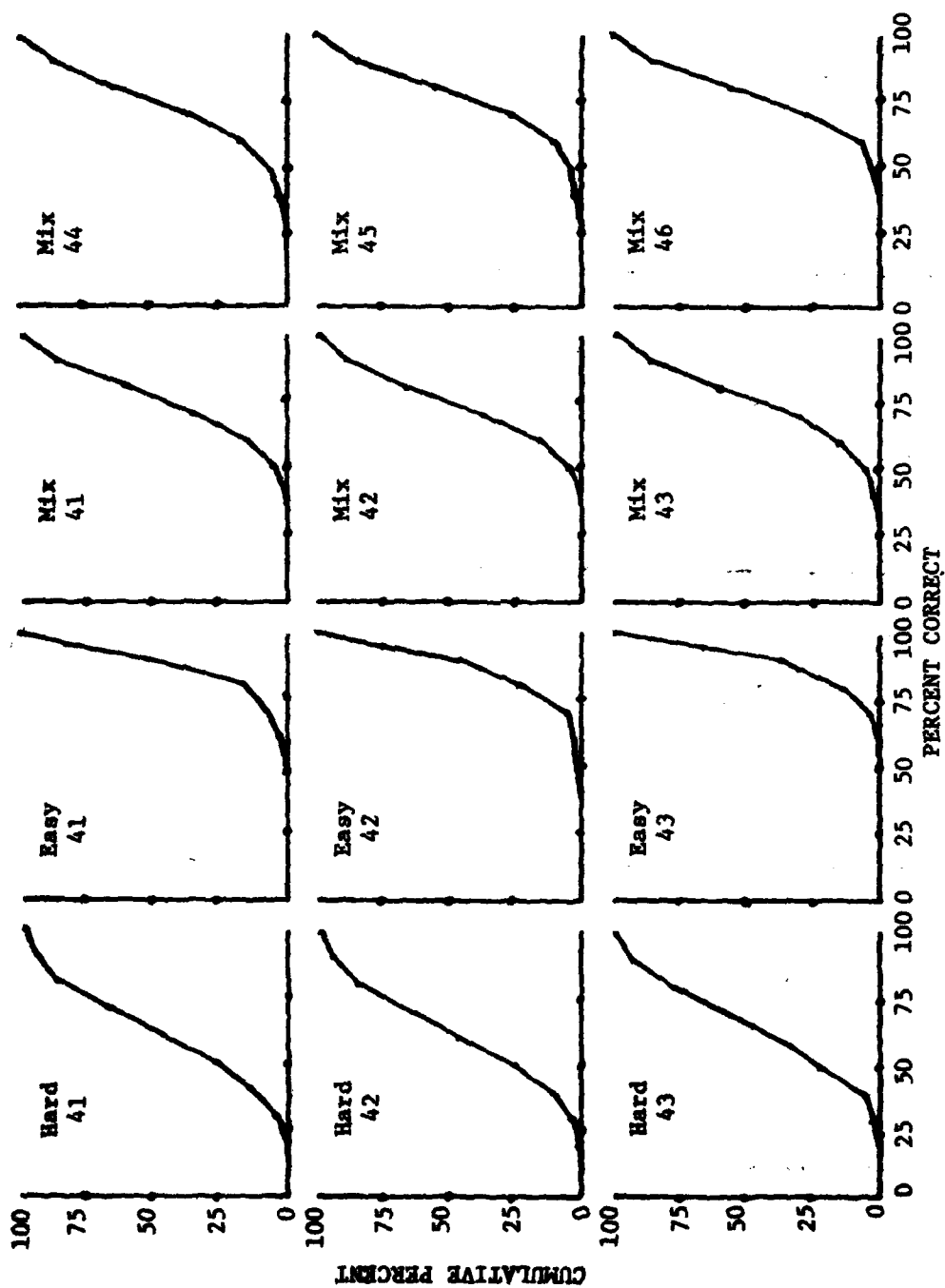


Figure 7 (cont): 40 Round Hard, Easy, and Mix Subtests

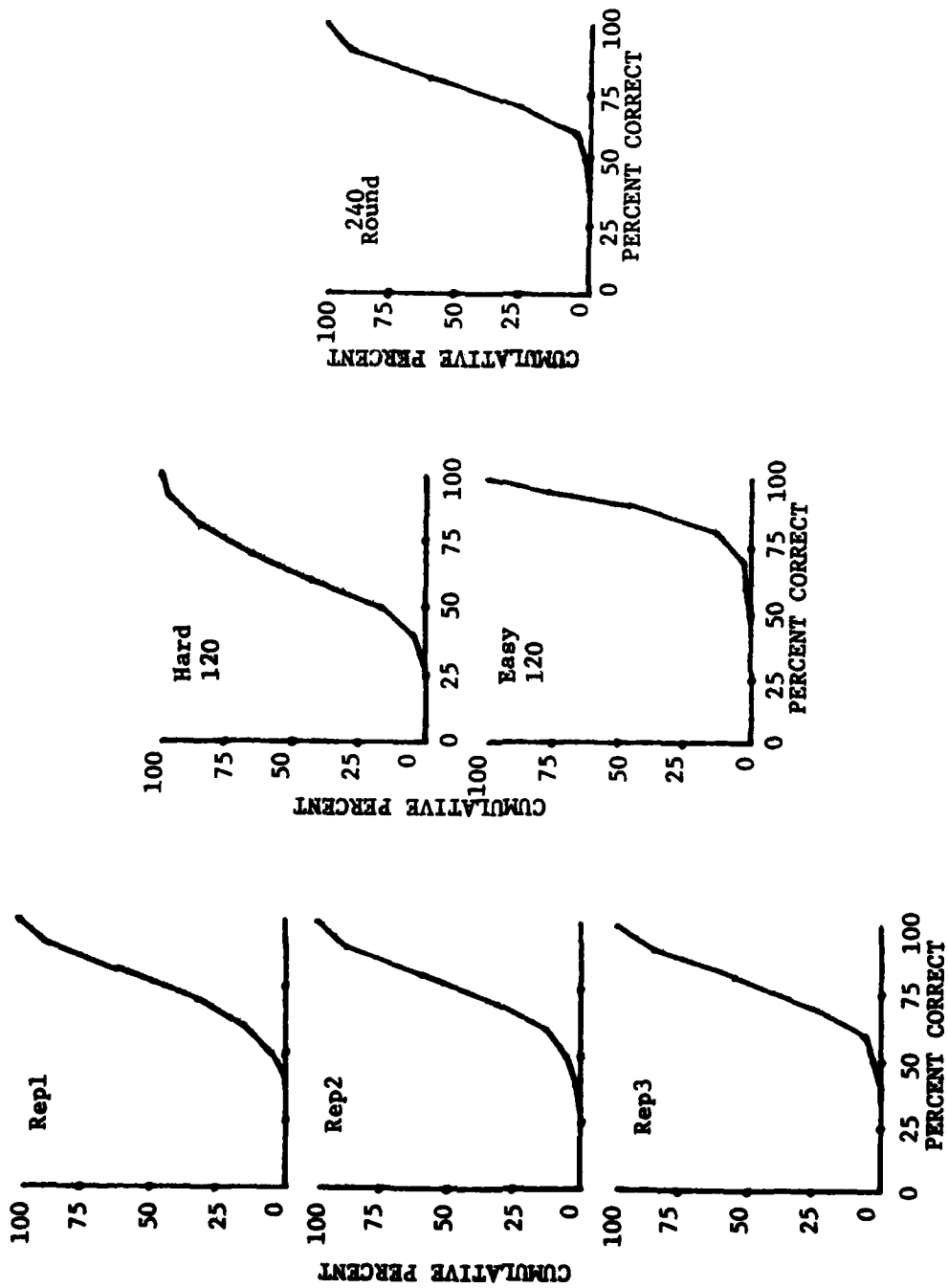


Figure 7 (cont): 80 Round Subtests, 120 Round Hard and Easy Criterion Tests, 240 Round Criterion Test

The ANOVA summary table for the MPFQC performance data is found in Table 7. Table 7 also includes the proportions of total variance accounted for by each of the main effects and interactions. All of the main effects and most of the interactions were statistically significant at the  $\alpha = .05$  level or beyond. These ANOVA results suggest that the MPFQC is an unstable instrument producing results that would be difficult to generalize or interpret.

The results for the proportion of variance accounted for suggest a different interpretation. This difference in possible interpretation highlights the importance of carefully considering the meaning of statistical significance, particularly when dealing with extremely powerful tests. The largest source of variance is the error term, accounting for 39% of the total variance. An additional 25% of the variance is accounted for by differences in the tables, and 10% more by individual differences between the examinees. These three factors account for 74% of the total variance. The other two main effects account for less than 1% of the variance, and the interactions which have more than 1% of the variance associated with them all include either Persons or Tables or both among the interacting factors. These results suggest that the MPFQC is relatively stable across groups of shots and test repetitions but that the homogeneity of the performance required by the different tables is questionable. These results have the same general implications as those for the descriptive data and seem to be more reasonable than the more extreme ANOVA F-ratio results.

Since all of the results suggest that differences between tables are important, two further statistical tests were performed. The mean

Source of Variance	df	M.S.	Quasi F-ratio	Adjusted df Numerator	Adjusted df Denominator	Proportion of Variance
P (Persons)	236	12.80	3.93****	272	705	.1027
S (Score Groups)	1	7.70	5.96**	1	25	.0006
T (Tables)	7	732.71	79.11****	7	142	.2454
R (Repetitions)	2	34.75	12.55****	2	87	.0041
PS	236	1.05	1.09	236	472	.0017
PT	1652	1.90	1.33****	3100	4695	.0536
PR	472	2.45	2.52****	472	472	.0444
ST	7	2.26	1.94*	13	75	.0007
SR	2	.40	.41	2	472	0
TR	14	4.31	2.82***	4	99	.0032
PST	1652	.91	1.11	1652	3304	.0144
PSR	472	.97	untestable			.0582
PTR	3304	1.14	1.38****	3304	3304	.0769
STR	14	.68	.83	14	3304	0
PSTR	3304					.3939

\*p < .05

\*\*p < .025

\*\*\*p < .01

\*\*\*\*p < .001

Table 7: Analysis of Variance Summary Table and Proportion of Total Variance Accounted For by Main Effects and Interactions  
(Completely Crossed, Mixed Model: P, S, R Random, T Fixed)

score for the hard tables was compared to the mean score for the easy tables using the Tukey test (Winer, 1971). The means are statistically significantly different ( $Q(8,42) = 36.93, p < .001$ ). The Tukey test was also used to show that the mean score of the easiest hard table is statistically significantly different from the mean score of the hardest easy table ( $Q(8,42) = 23.24, p < .001$ ). These results further support the notion of a two part domain to describe the MPFQC.

#### Comparison of the Scoring Models: 240 Round Criterion

The models were compared on the basis of their recommended criterion scores, the misclassification rates observed when subtest decisions based on each model's recommended criterion score were compared to the decisions based on the full 240 round test, the difference between the observed misclassification rates and the misclassification rates that are predicted by the statistical properties of each model, and the accuracy of the models' subtest true score estimates compared to the true score defined by the 240 round test. Table A summarizes the results for the recommended criterion scores and the misclassification rates. Table B summarizes the results of the true score estimations. Table A also includes entries for an empirical best criterion score. The empirical best criterion score was defined as that score which resulted in the lowest total observed misclassification rate, where total misclassification equals the sum of the false positive rate and the false negative rate. Each factor considered in the comparison of the models will be treated in turn.

#### Criterion Score

Since the proportion correct and Bayesian models do not use ob-

served test data in their procedures for determining a criterion test score, the criterion scores were constant for each test length. The binomial error model procedure could suggest a different criterion score for each testing occasion since it is dependent on the distributions of the observed scores. The empirical best criterion score can also vary for different testing occasions depending on the distribution of observed scores.

For the 10 round subtests, the proportion correct and Bayesian models' recommended criterion scores are 7 and 8 correct. The reason for two scores is that there was no rationale for deciding that one score was clearly more advantageous than the other. The lower score is slightly more favorable if false negative errors are critical and must be kept to a minimum, the higher score is slightly more favorable if false positive errors are to be minimized. Multiple criterion scores are also recommended for the longer subtests by the proportion correct and Bayesian models for the same reason.

The recommended 10 round criterion scores for the binomial error model vary from 4 to 8. In the case of the hard subtests (Tables 11 - 14, 21 - 24, 31 - 34), the criterion score is 8 for all but one test which has a criterion score of 7. For the easy tests (Tables 15 - 18, 25 - 28, 35 - 38), the criterion scores vary from 4 to 7. The lower criterion scores observed for the easy tests never resulted in lower total misclassification than a criterion score of 7 and usually resulted in a higher total misclassification rate.

The empirical best criterion scores for the 10 round subtests vary from 3 to 9. For the hard subtests, the criterion scores vary

from 3 to 6, always lower than those suggested by the models. For the easy subtests, a criterion score of 9 is empirically best in two cases. For the other ten easy subtests, the empirical best criterion score is either 7 or 8.

For the 20 round subtests, the proportion correct and Bayesian models' recommended criterion scores are 14 and 15. Test difficulty was, again, an important factor in the criterion score recommended by the binomial error model and the empirical best procedure. For the 20 round hard subtests, the binomial error model criterion scores are 14 or 15, for the easy subtests the criterion scores vary from 11 to 13, and for the mix subtests of intermediate difficulty, criterion scores of 13 or 14 are recommended. The lower easy subtest criterion scores did not lower the total misclassification rate and usually increased it relative to the other models. However, in two cases for the mix subtests, a criterion score of 13 did result in lower total misclassification than criterion scores of 14 or 15. The empirical best criterion scores for the hard subtests are lower than those recommended by the models, varying from 8 to 10. For the easy subtests, the empirical best criterion scores are 16 or 17, in all cases higher than the models' criterion scores. For the mix subtests, the empirical best criterion scores are closer to the models', varying from 12 to 14.

The 40 round subtest results show similar trends. The proportion correct model's criterion scores are 27, 28, and 29. The Bayesian model's criterion scores are 28 and 29. For the 40 round hard subtests, the binomial error model's criterion scores are all 29, for the easy subtests, they vary from 25 to 27, and for the mix subtests, criterion

scores of 27 or 28 are suggested. The lower criterion scores recommended for the easy subtests did not result in decreased total misclassification. The empirical best criterion scores for the 40 round hard subtests vary from 18 to 24, for the easy subtests they are 33 or 35, and for the mix subtests they are 27 in five cases and 26 in the one remaining case.

The 80 round subtests closely resembled the mix subtests in their test characteristics. The criterion scores reflect this similarity in their homogeneity across procedures. The proportion correct model's criterion scores are 54, 55 and 56. The Bayesian model's criterion scores are 56 and 57. Binomial error model criterion scores are 55 or 56 and the empirical best criterion scores are 53, 54, or 56.

Overall, the criterion scores recommended by the models are remarkably similar to each other and close to what one would choose on purely intuitive grounds. That is, given that the Military Police School uses a criterion score of 70% hits for qualification, the criterion scores would be 7, 14, 28, and 56 for the 10, 20, 40, and 80 round subtests, respectively. The only differences in recommended criterion scores among the models occurred with the binomial error model which tended to suggest lower scores for the easy subtests. These results for the binomial error model reflect its use of empirical data in determining a criterion score. However, the binomial error model's equation for estimating true scores appears to have overestimated true scores for the easy subtests to such an extent that decision making accuracy suffered.

Both the models' recommended criterion scores and the intuitively appealing 70% hits criterion are relatively poor choices compared to the



empirical best criterion scores. For the hard subtests, the empirical best criterion scores tended to be lower than the criterion scores suggested by the models. Two factors help explain this result. First, the nonmaster group tended to get very low scores on the hard subtests. The masters, on the other hand, were able to achieve at least moderate scores on the tests. Therefore, for low criterion scores, relatively few masters were misclassified and most of the nonmasters were correctly failed. As the criterion score was raised, relatively little gain in reducing the false positive rate occurred but the false negative rate climbed rapidly. The second factor contributing to lower empirical best criterion scores on the hard subtests is the distribution of masters and nonmasters in the examinee group. Based on the full 240 round test, 61 persons (25.7%) were nonmasters and 176 persons (74.3%) were masters. Therefore, the maximum false positive misclassification rate (which would occur for criterion score equals 0; all pass) is .257. Since no false negative misclassifications can occur at that criterion score, the total misclassification will also be .257. The maximum false negative rate (at criterion score equals 11; all fail) is .743. Since no false positives can occur when all persons fail, the total misclassification rate will also equal .743. Since the false positive rate has a lower limiting value than the false negative rate by a considerable amount, total misclassification will tend to be lower for cases where the false positive rate is high relative to the false negative rate than for cases where the false positive rate is low relative to the false negative rate. For the hard subtests a relatively low criterion score correctly failed a large proportion of the nonmasters while cor-

rectly passing most of the masters. By increasing the criterion score, better decisions for the nonmasters can be achieved but only by sacrificing some of the correct classifications for masters. Since there are far more masters than nonmasters in the examinee group, the best strategy is to choose a lower criterion score if the goal is to keep total misclassification to a minimum.

For the easy subtests, the models' suggested criterion scores tended to be close to or equal to the empirical best scores. In contrast to the hard subtests, when the empirical best criterion scores did differ from the models', they tended to be higher. All persons, regardless of their status on the 240 round test, tended to get high scores on the easy subtests. Therefore, a relatively high criterion score was needed to fail most of the nonmasters. Since masters tended to get high scores, the cost in a high false negative rate that was found for the hard subtests with relatively high criterion scores was not the case for the easy subtests. Rather, the false negative rate remained low while the false positive rate tended to approach its limiting value of .257.

#### Observed Misclassification Rates

The observed misclassification rates were defined as the proportions of all classifications that were false positives or false negatives. These were summed to yield total misclassification rates. In addition, the ratios of the false positive to the false negative rates were computed. Since the applications of the models were predicated on an equal loss ratio of false positive to false negative errors, desirable values for the false positive to false negative (FP:FN) ratio are close to 1.0.

The overall impression of the observed misclassification rates is that they are high for the models and the empirical best approach regardless of test length. Figure 8 shows each model's average observed misclassification rate for each test length. The results show a clear pattern in the false positive and false negative rates with respect to test difficulty. In the case of the hard subtests, the false positive rates are relatively low and tend to decrease with increasing test length. The false negative rates are high and show little change as the test length increases. Total misclassification rates for the hard subtests decrease slightly with increasing test length. In the case of the easy subtests, these results are reversed. False positive rates are relatively high and tend to increase slightly with increasing test length. False negative rates are very low and decrease for the longer subtests. There is little change in the total misclassification rates as a function of test length. The moderately difficult mix subtests produced the most reasonable results. Both the false positive and false negative rates are relatively low, they are comparable to one another, and they decrease with increasing test length. The total misclassification rates decrease as a function of increasing test length. For all test lengths the total misclassification rates are lowest for the mix subtests and highest for the hard subtests.

The misclassification rates for the empirical best criterion scores show a different pattern than those of the models. In all cases, the false positive rates are higher than the false negative rates. The false positive and total rates decrease for the longer tests, but the false negative rates remain relatively constant with

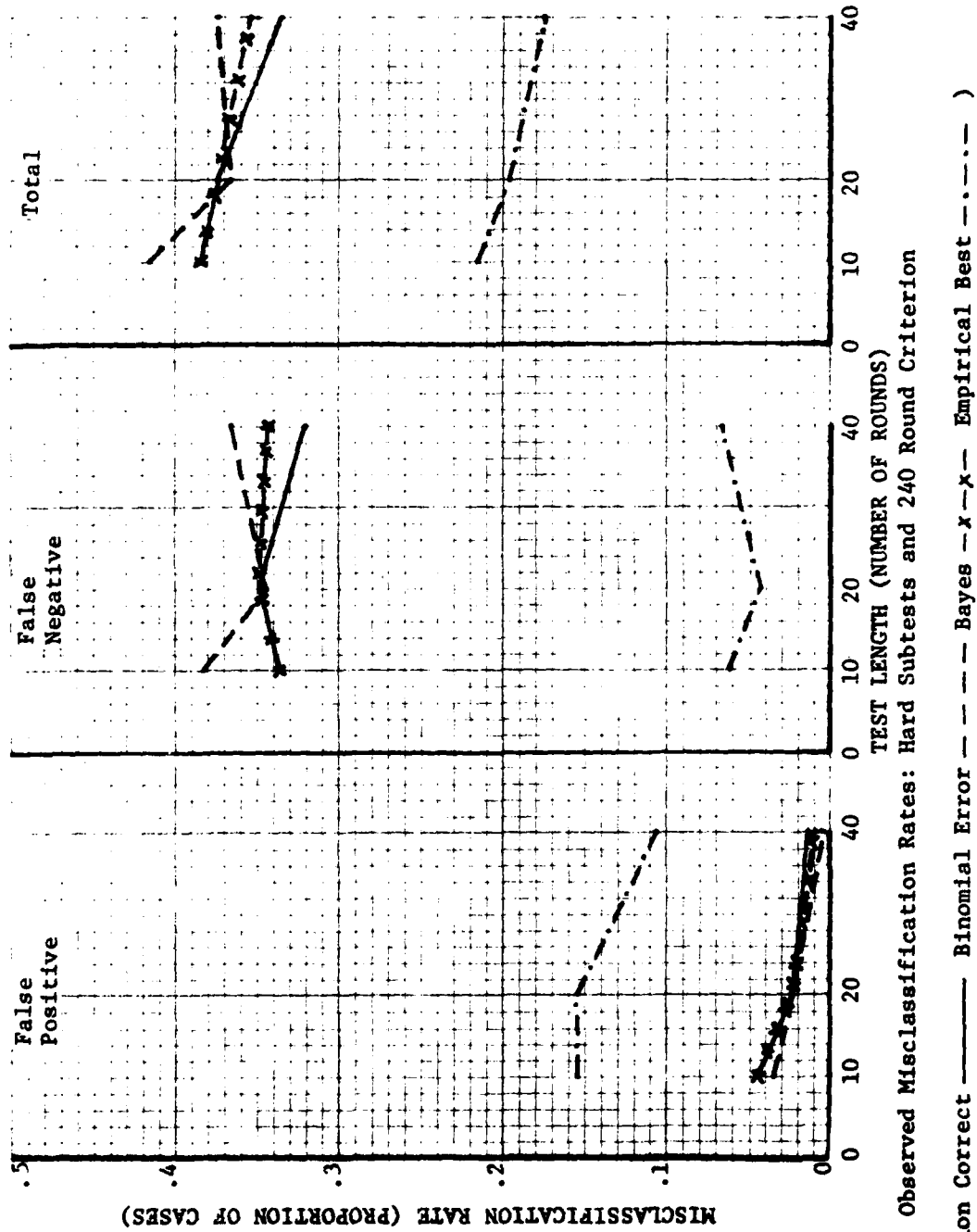


Figure 8: Observed Misclassification Rates: Hard Subtests and 240 Round Criterion

(Proportion Correct — Binomial Error - - - Bayes - x - x - Empirical Best - . . . )

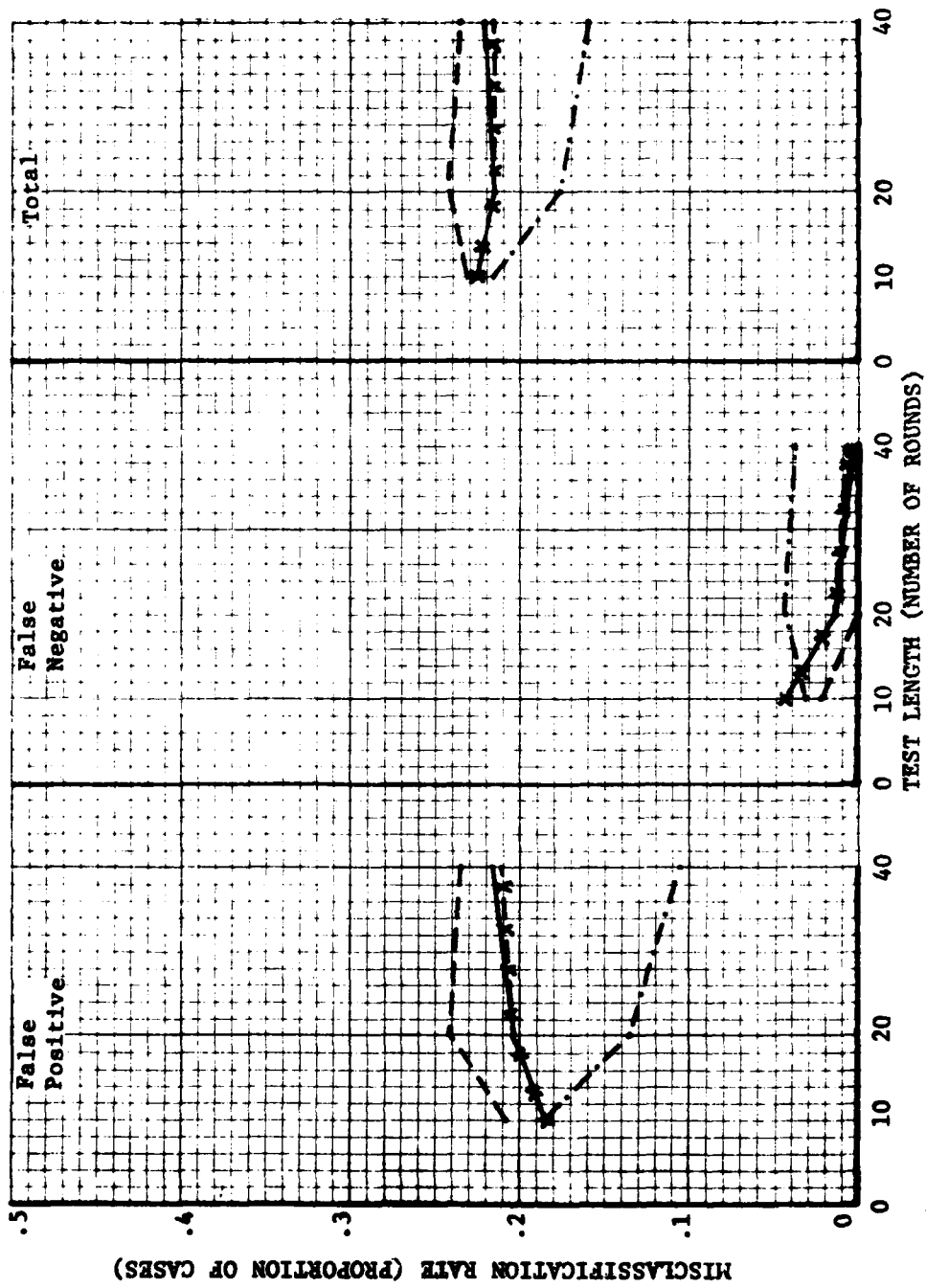


Figure 8 (cont): Easy Subtests and 240 Round Criterion

(Proportion Correct ——— Binomial Error --- Bayes -x-x- Empirical Best -.-.- )

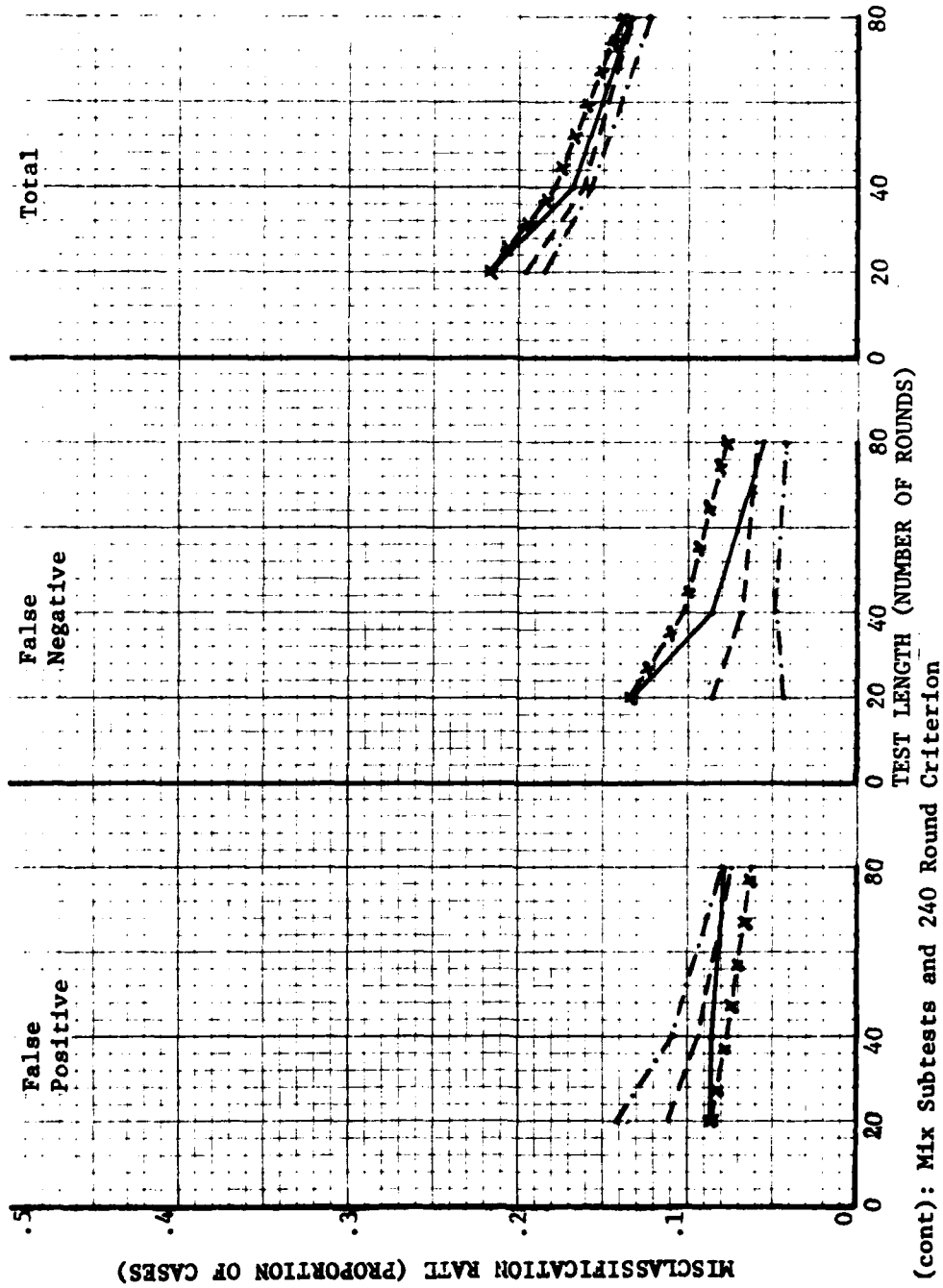


Figure 8 (cont): Mix Subtests and 240 Round Criterion

(Proportion Correct ——— Binomial Error — — — Bayes — x — x — Empirical Best — · — · — )

test length. There is no advantage in lower total misclassification shown by the hard, easy, or mix tests. Although total misclassification is always lower for the empirical best procedure than for the models, this is not the case for the false positive and false negative rates. False positive rates are higher for the empirical best procedure for the hard and mix subtests, and the false negative rates are higher for the easy subtests. The empirical best procedure achieved the lowest total misclassification rates because neither the false positive nor the false negative rates took on extremely high values. Rather, moderate misclassification rates are generally the case for this procedure.

The results for the FP:FN ratios show patterns similar to those for the observed misclassification rates. In all cases, the FP:FN ratios tend to be considerably different than 1.0 and show relatively little improvement with test length. The least desirable results were generally obtained for the hard subtests and the most desirable results were found for the mix subtests. An exception to this finding was that for some of the easier subtests, the false negative rate fell to 0, leading to undefined values for the FP:FN ratios. In general, the FP:FN ratios obtained with the empirical best criterion score are closer to 1.0 than those obtained with the models' criterion scores.

The analysis of the observed misclassification rates highlights the similarities among the models rather than any differences between them. There were very few differences between the models in the criterion scores suggested. Where differences did appear they did not tend to improve either classification accuracy or the FP:FN ratio.

None of the models compared well to the empirical best results. What appeared to be the most important criterion in determining a best criterion score was the test difficulty. Figures 9 and 10 show this relationship. The data are the test difficulties and the best criterion scores, expressed as a percent correct, for all test lengths. For best criterion score defined as that with the lowest total misclassification (Figure 9), the correlation was .887. For best criterion score defined as that with the FP:FN ratio closest to 1.0 (Figure 10), the correlation was .882.

#### Expected Misclassification Rates

The expected misclassification rate data are summarized in Figure 11. The expected misclassification rates for the empirical best procedure were computed in the same way as for the proportion correct model using the empirical best criterion scores in place of the proportion correct criterion scores. The values for the expected misclassification rates are comparable theoretically, all being algebraically equivalent to the probability of being a master and failing (false negative) or the probability of being a nonmaster and passing (false positive). However, the data used to compute these values for the proportion correct model and the empirical best procedure are different than those used for the binomial error and Bayesian models. The expected misclassification rates for the proportion correct model and empirical best procedures represent what would be expected given the distribution of abilities according to the 240 round criterion test, a criterion score, and the properties of the binomial distribution. In the case of the binomial error and Bayesian models, the expected misclassification



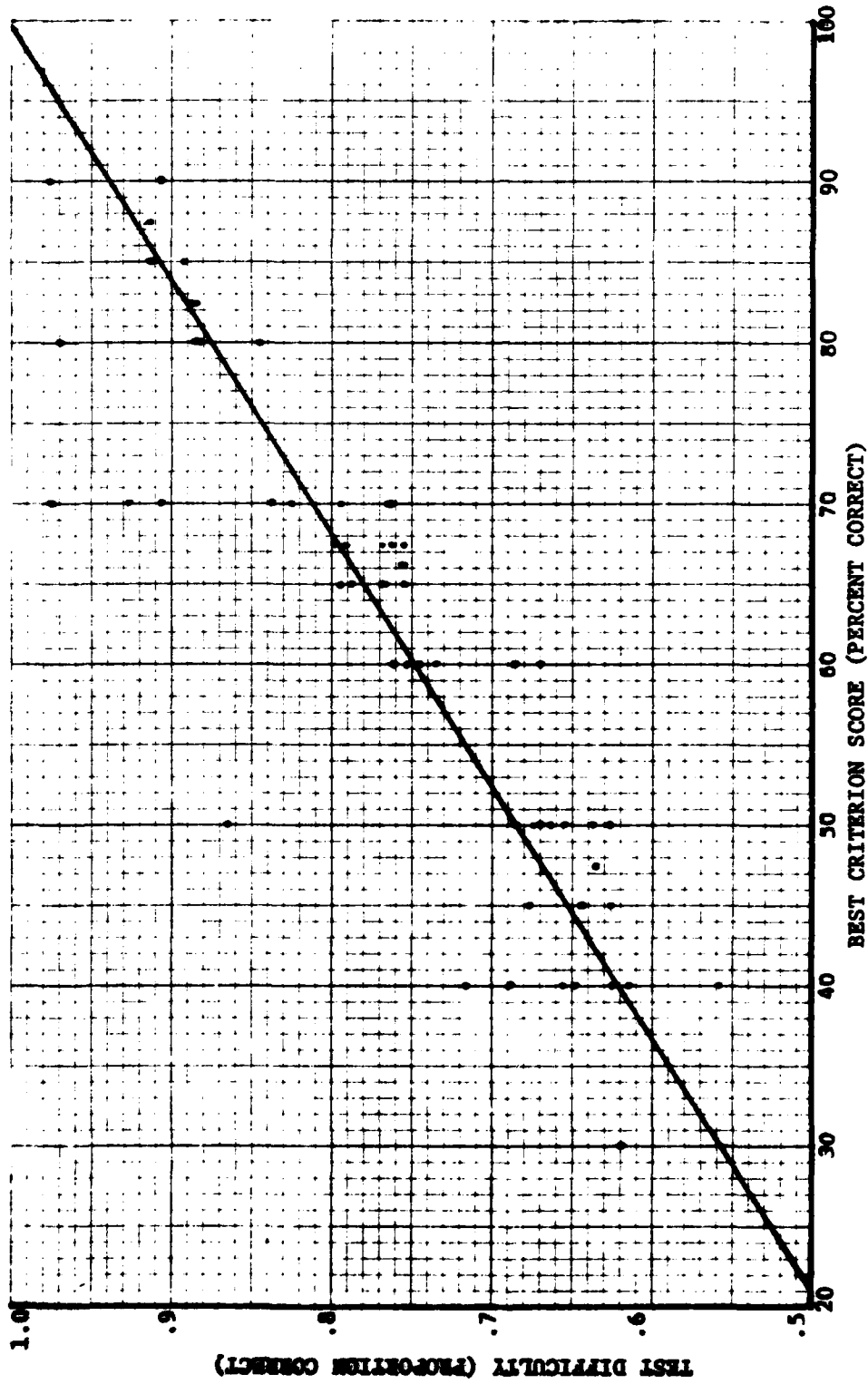


Figure 9: Scatterplot of Test Difficulty and Best Criterion Score When Best Criterion Score is Defined as the Score Producing the Lowest Total Misclassification Compared to the 240 Round Criterion

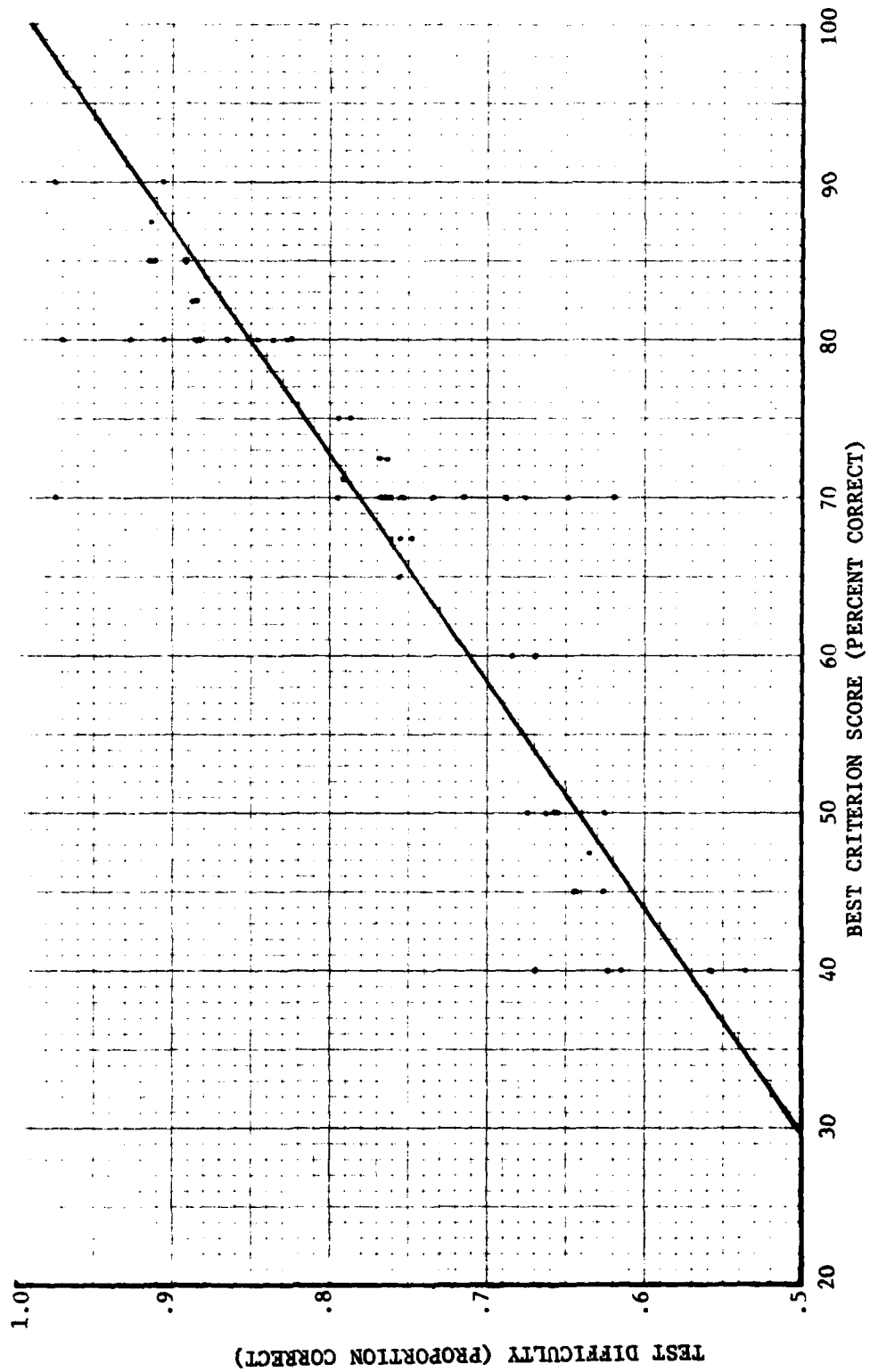


Figure 10: Scatterplot of Test Difficulty and Best Criterion Score When Best Criterion Score is Defined as the Score Producing the False Positive to False Negative Misclassification Rate Ratio Closest to 1.0 Based on 240 Round Criterion

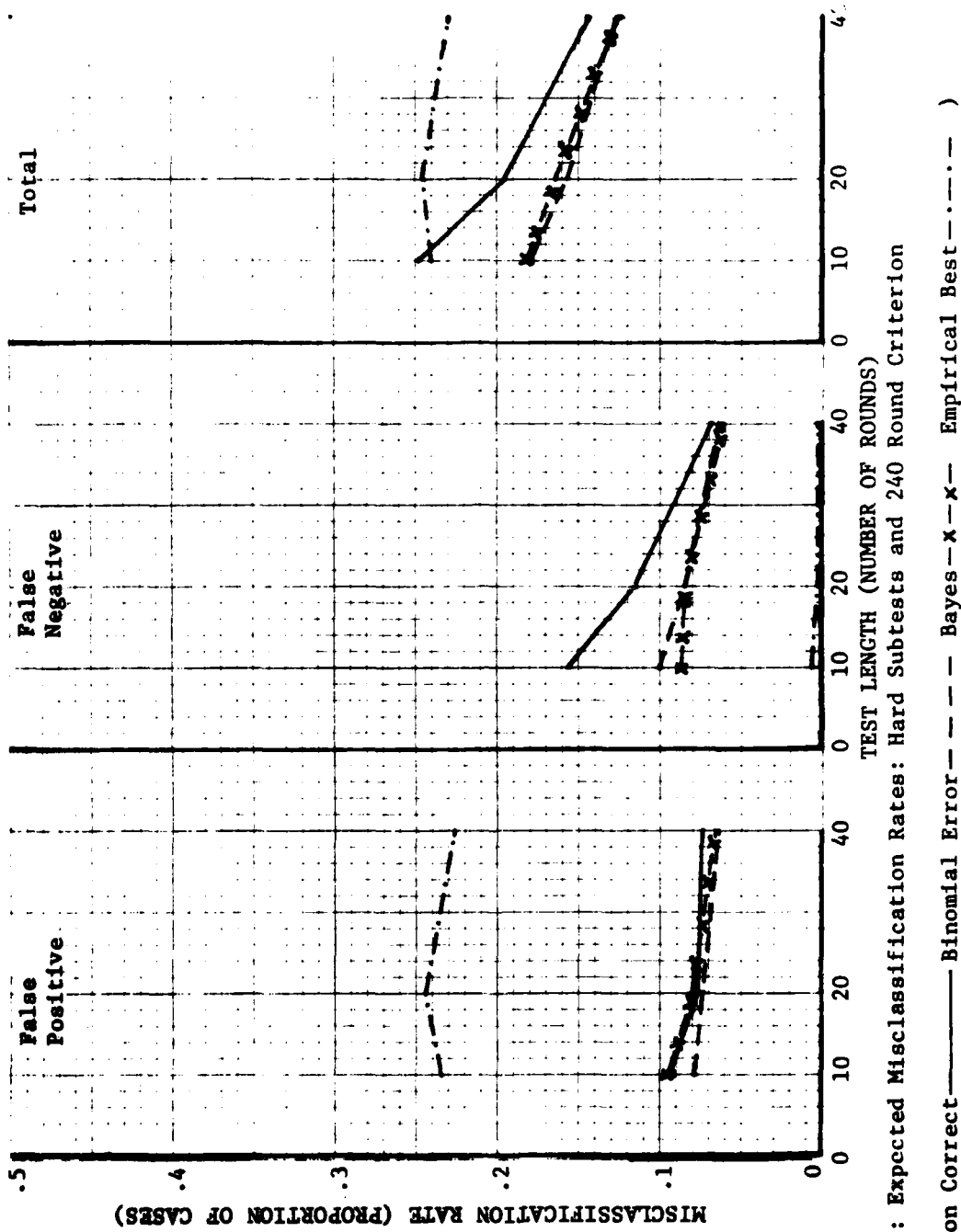


Figure 11: Expected Misclassification Rates: Hard Subtests and 240 Round Criterion

(Proportion Correct ——— Binomial Error - - - Bayes - x - x - Empirical Best - . - . - )

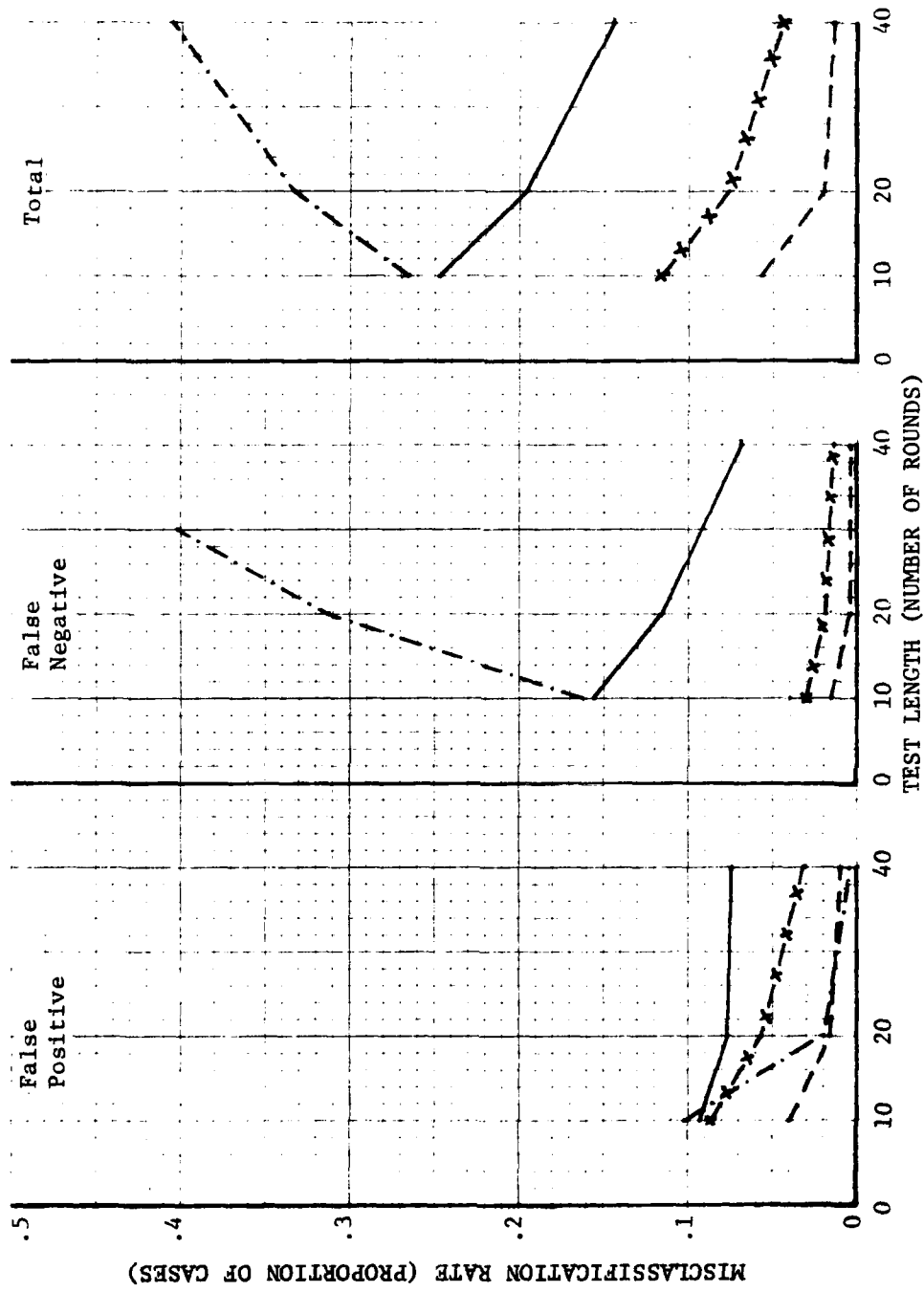


Figure 11 (cont): Easy Subtests and 240 Round Criterion

(Proportion Correct ——— Binomial Error - - - Bayes - x - x - Empirical Best - . - . - )

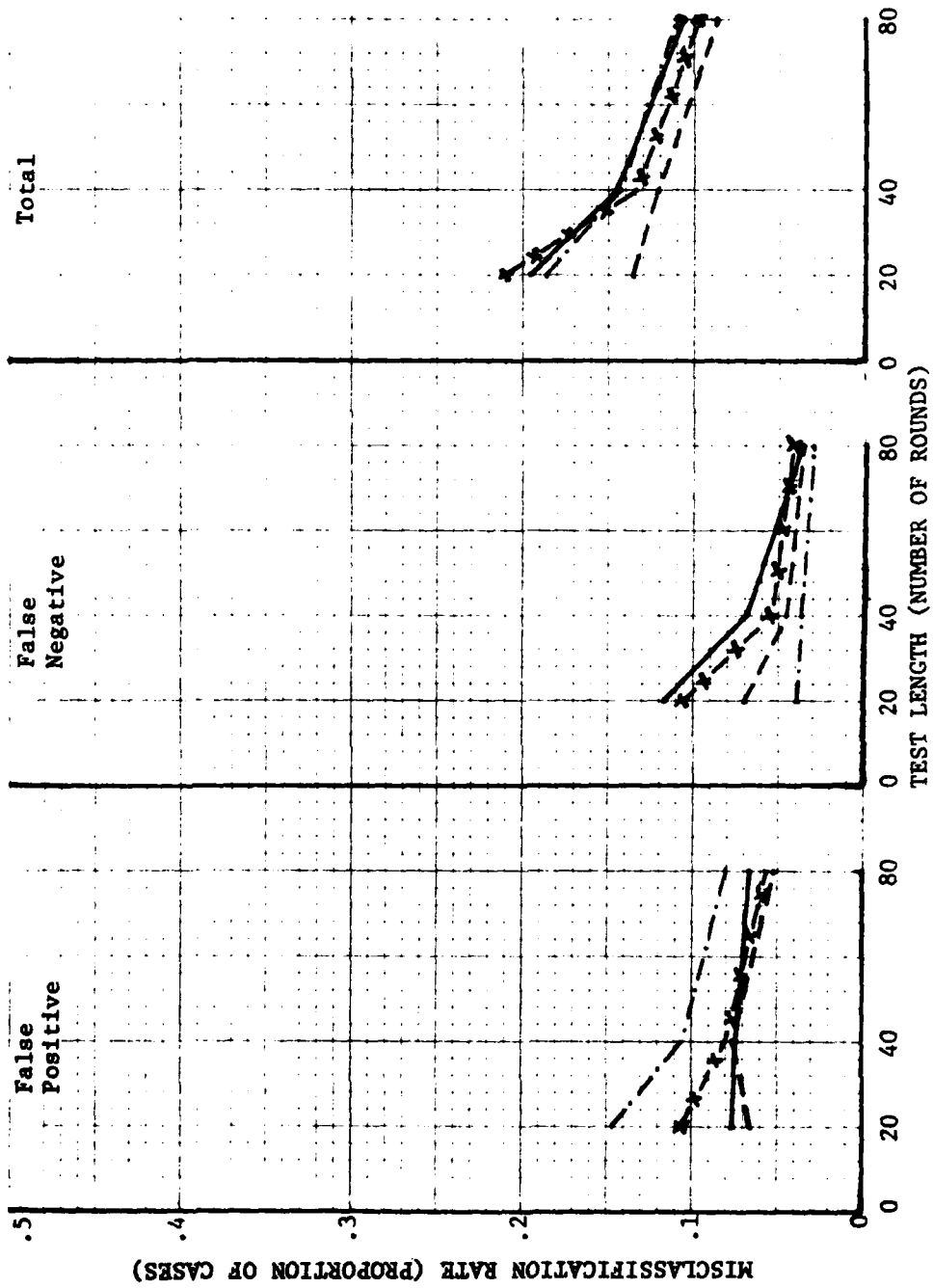


Figure 11 (cont): Mix Subtests and 240 Round Criterion

(Proportion Correct ——— Binomial Error — — — — Bayes — x — x — Empirical Best — · — · — )

tion rates are what would be expected given the distribution of observed scores, a criterion score, and the properties of the models. In other words, the expected misclassification rates for the proportion correct model and the empirical best procedure should be more sensitive to the 240 round true ability distribution, while the binomial error and Bayesian models' expected misclassification rates should be more sensitive to the observed score distributions for each subtest.

Since the proportion correct model's criterion scores were the same for the hard, easy, and mix subtests and the expected misclassification rates were computed on the basis of the constant 240 round test ability distribution, the expected misclassification rates are the same for all three types of tests. The false positive expected misclassification rate drops rapidly from the 10 to the 20 round subtests, and then declines more gradually through the 40 round subtests to the 80 round subtest results. The false negative rates fall more sharply across the test lengths. The curve for the total expected misclassification rates resembles the false negative curve.

These results illustrate the interaction between the binomial probability model and the distribution of masters and nonmasters in the examinee group. The average proportion correct score for the nonmasters on the 240 round test was .633, for masters it was .817. If one computes false positive and false negative rates for these values without adjusting for the relative proportions of masters and nonmasters in the group, the curves look similar to those actually obtained. However, the false negative rate is always less than the false positive rate. The fact that there are approximately three times as

many masters in the group causes the expected false negative rates for these data to exceed the false positive rates for the 10 round and 20 round subtests. In the case of the 40 round and 80 round subtests, the probability of a master failing gets sufficiently low that despite the disproportionate number of masters in the group, the expected false negative rates fall below the expected false positive rates.

The large differences between the expected misclassification rates for the empirical best criterion scores and the proportion correct model scores are indicative of the differences between the subtests and the 240 round criterion test. In the case of the hard subtests, relatively low criterion scores were required to produce the empirical best total observed misclassifications. These low scores produced high expected false positive and extremely low expected false negative rates. In the case of the easy subtests, these results were reversed. The relatively high criterion scores which were necessary to produce the empirical best total observed misclassification rates produced low false positive expected rates and very high expected false negative rates when applied to the 240 round test ability distribution. The expected misclassification rates for the moderately difficult mix subtests, which closely resembled the 240 round test with respect to mean scores and other test characteristics, are much closer to the proportion correct curves.

The expected misclassification rates for the binomial error and Bayesian models can also be understood in terms of the observed score distributions for the different difficulty type tests. In the case of the hard subtests, the median scores tended to be about 65% correct.

This is slightly below most of the criterion scores suggested by the models. Therefore, the probability of observing scores at or above criterion was slightly less than the probability of observing scores below criterion. Since the expected false positive rate is proportional to the probability of observing scores at or above criterion and the expected false negative rate is proportional to the probability of observing a score below criterion, one would expect the false positive rate to be slightly lower than the false negative rate. That was, in fact, the case. The same general line of reasoning explains the misclassification rates for the easy subtests. The median for the easy subtests tended to be about 90% correct. The probability that an observed score at least equaled the criterion score was greater than the probability of observing a score less than the criterion. Hence, one would expect, and the results show, the false positive rates to be greater than the false negative rates. The differences in rates between the binomial error and Bayesian models are due to differences in the probability distributions which describe the probability of an individual being a master or nonmaster given the observed score. For the Bayesian model these distributions are based on a prior distribution which is common to all observed scores. For the binomial error model, these distributions reflect the observed score distribution of each subtest.

For the mix subtests, the median scores tended to be about 75%. Since this is slightly above most of the criterion scores, the probability that a score was above criterion was slightly higher than the probability that the score was below criterion. The implication, and



what was observed, is that the false positive rate should be slightly higher than the false negative rate. The curves for the binomial error and Bayesian models were much more similar to the proportion correct curves for the mix subtests than for the hard or easy subtests, reflecting the similarity of the mix subtests to the 240 round criterion test.

#### Observed versus Expected Misclassification Rates

The difference between observed and expected misclassification rates is one of the most important criteria on which to compare the models. This is simply because it is more advantageous to employ a scoring procedure that produces predictable results than one which does not. If a statistical model represents the phenomena underlying the data, then one would expect the model to produce predictable results, and the differences between observed and expected misclassification rates should be small. Such small differences would be expected regardless of the extent of the observed misclassification. The results, expressed as averages of the absolute values of the differences between observed and expected misclassification rates are summarized in Figure 12.

For the hard subtests, all of the models and the empirical best procedure overestimated the false positive misclassification rates. The models' results are nearly identical and are indicative of relatively accurate predictions. The empirical best procedure was much less accurate and shows a steady and appreciable increase in overestimation with increasing test length. The hard subtests' observed versus expected difference results for false negative and total misclassifi-

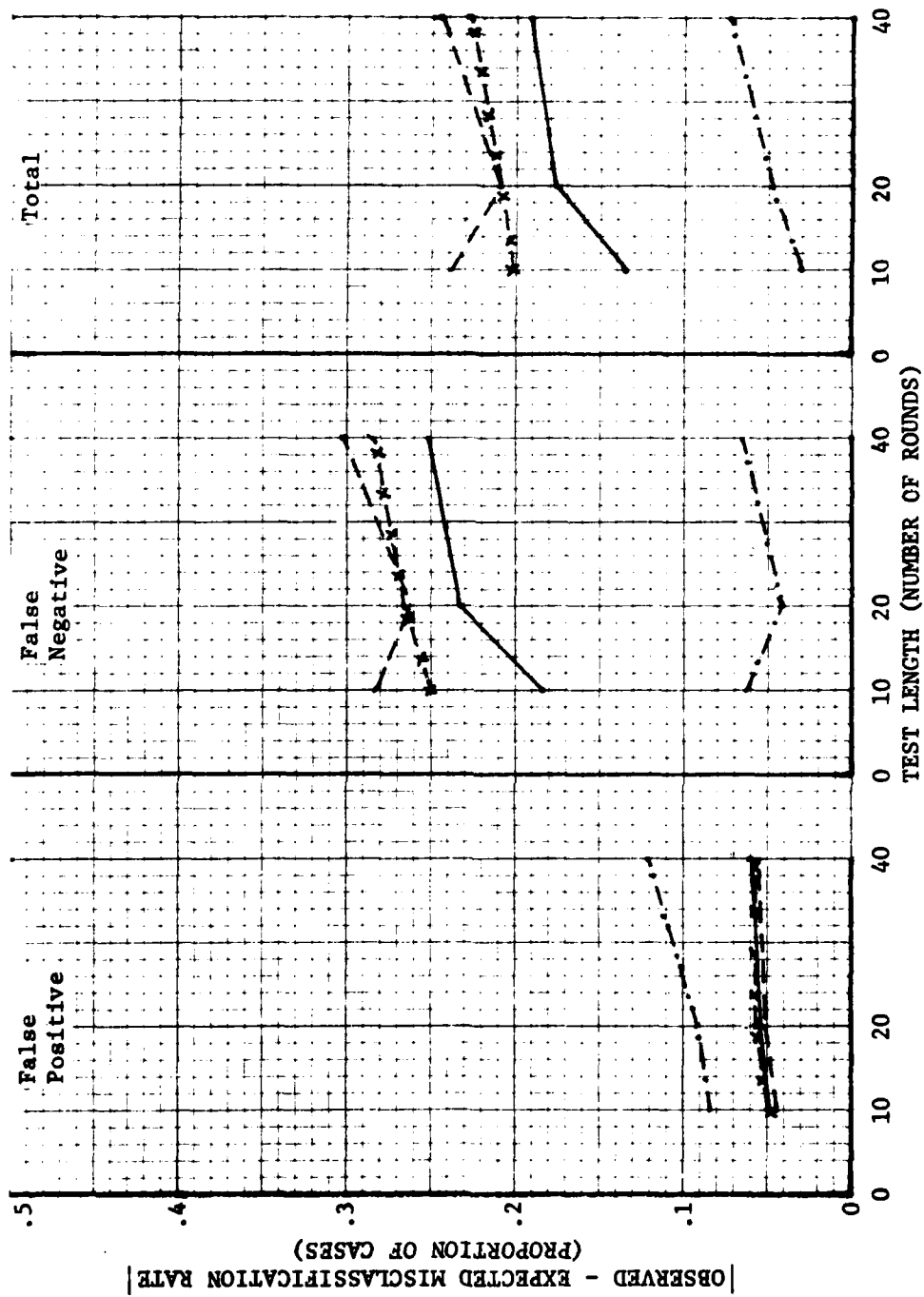


Figure 12: Absolute Values of Differences Between Observed and Expected Misclassification Rates:  
 Hard Subtests and 240 Round Criterion  
 (Proportion Correct — Binomial Error — Bayes — Empirical Best — )

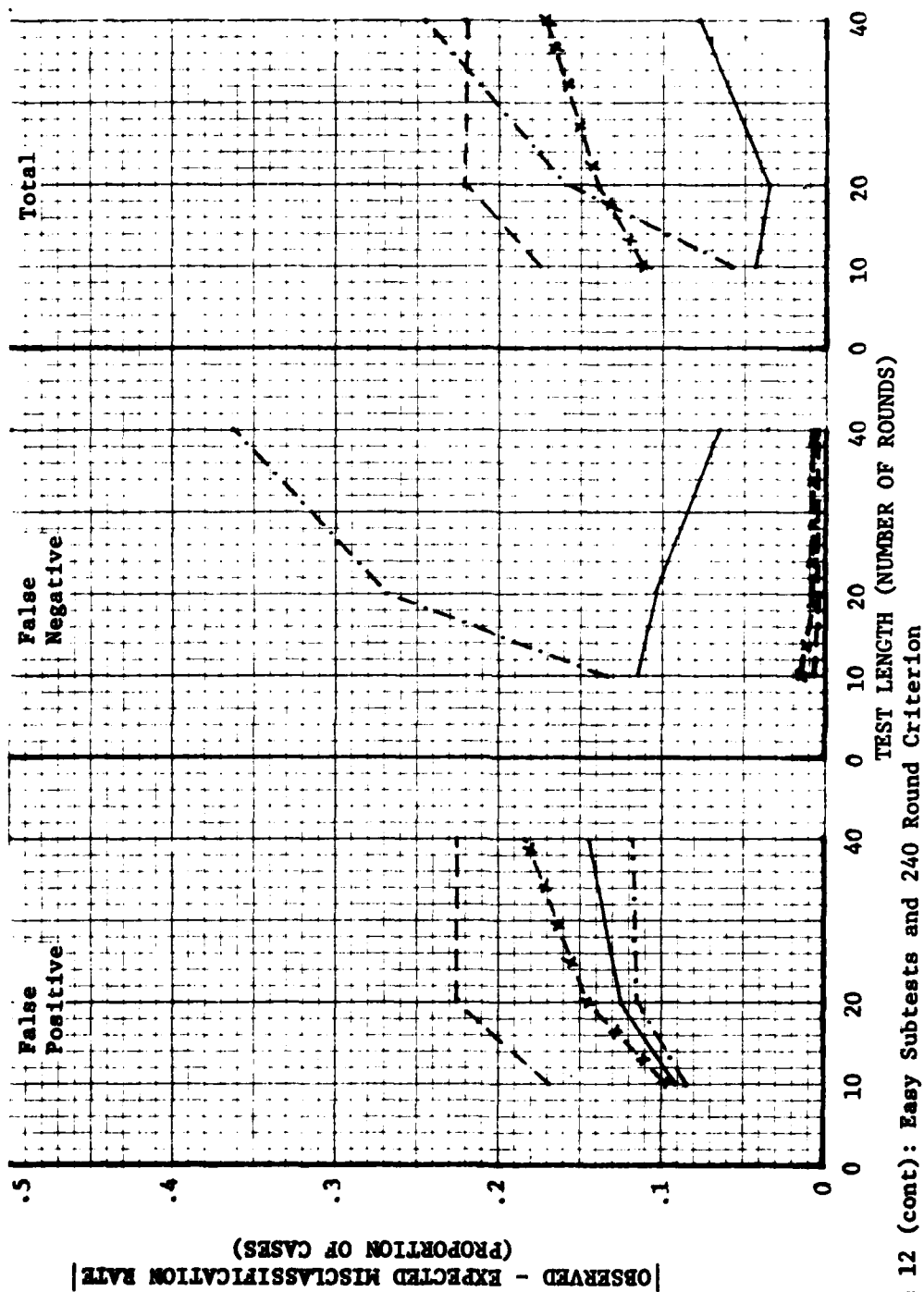


Figure 12 (cont): Easy Subtests and 240 Round Criterion

(Proportion Correct — Binomial Error — — — Bayes — x — x — Empirical Best — · — · )

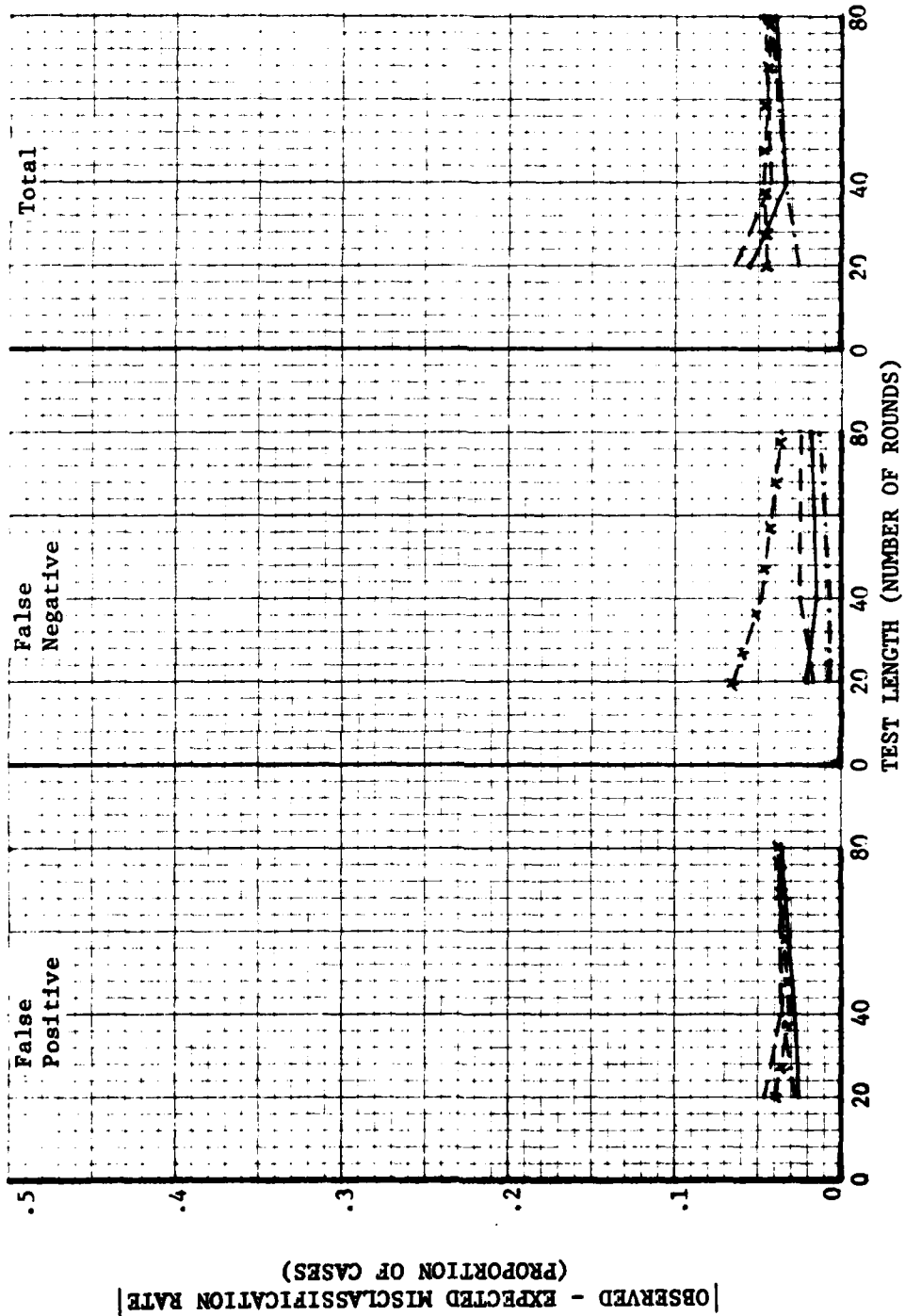


Figure 12 (cont): Mix Subtests and 240 Round Criterion

(Proportion Correct ——— Binomial Error — — — — Bayes — x — — Empirical Best — · — · — )

cation rates are the reverse of the false positive results. The empirical best procedure shows relatively small differences and the three models' differences are high. All of the procedures nearly always underestimated the observed false negative misclassification rates. However, the models nearly always underestimated the total observed misclassification, while the empirical best procedure nearly always overestimated it. For both the false negative and total difference data, the proportion correct model performed better than the binomial error or Bayesian models.

The easy subtests' results are much more varied than those for the hard subtests. In the case of the false positive errors, all of the procedures nearly always underestimated the observed error. The empirical best procedure produced the most predictable results, followed by the proportion correct model. The binomial error model performed most poorly. In the case of the false positive errors, the results are reversed. All of the procedures nearly always overestimated the observed error. The empirical best procedure performed very poorly. The proportion correct model's results show relatively good prediction and the binomial error and Bayesian models produced very good results. When the differences were summed to reflect total misclassification, the proportion correct model was best. The empirical best procedure did well for the 10 round subtests but its performance rapidly declined until it looked worst for the 40 round subtests. The binomial error model was less predictable than the Bayesian model. All three models tended to underestimate the total observed misclassification while the empirical best procedure tended to overestimate it.

The mix subtests' results are very encouraging. There is little difference between the results for any of the procedures, and all performed well. For the false positive errors the models overestimated the observed error slightly more often than they underestimated it. The empirical best procedure tended to overestimate the observed error. All of the procedures tended to underestimate the observed false negative misclassification rates. In the case of the total misclassification for the mix subtests, the models tended to underestimate the misclassification rates while the empirical best procedure tended to overestimate them.

These data emphasize the complexity of the interactions between the ability distribution as defined by the 240 round criterion test, the observed score distributions, and the data used to compute expected misclassification rates. The hard subtests were characterized by low scores for all examinees, relatively high criterion scores for the proportion correct, binomial error, and Bayesian models, and relatively low empirical best criterion scores. The high criterion scores recommended by the models led to low observed false positive rates and high false negative rates. In the case of the proportion correct model, the expected false positive rate was also moderately low, so that the differences between observed and expected false positive misclassifications were relatively small. The observed false negative rate was, however, much higher than what would have been expected by the 240 round criterion ability distribution. Therefore, the differences between observed and expected false negatives for the proportion correct model were large. Since the false positive rate was overestimated

and the false negative rate was underestimated, there was some tendency for the two effects to cancel each other in the differences between observed and expected total misclassifications. The differences for the false negative rates were, however, so large that the total differences were also large.

The results for the binomial error and Bayesian models have similar values to those for the proportion correct model but for different reasons. The low observed scores led to low expected false positive rates since the frequency of passing scores, which contribute to the false positive rate, was relatively small. Low observed scores can also imply relatively low expected false negative rates if the majority of scores are sufficiently low to insure that the portions of the ability distributions for failing scores which exceed the criterion ability (in this case .70) are small. The data illustrate these effects by the low expected false positive and false negative rates. The difference data show relatively good predictions of false positive misclassification rates, poor predictions of false negative misclassification rates, and, despite some canceling of the two types of error, large differences between observed and expected total misclassifications.

The empirical best criterion scores for the hard subtests were lower than those of the models. This led to higher observed false positive and lower observed false negative rates. When the low criterion scores were applied to the 240 round criterion ability distribution, high false positive and low false negative rates were predicted. This led to larger differences between the observed and expected false

positive rates than were the case with the models, but considerably smaller differences between observed and expected false negative and total misclassification rates.

The easy subtests were characterized by high scores, few differences in the models' criterion scores, and slightly higher empirical best criterion scores. Although the models' criterion scores were relatively high, they were not high enough to prevent high observed false positive rates, but the observed false negative rates were low. The proportion correct model produced moderate values for both the expected false positive and false negative rates. The expected false positive rate underestimated that observed to almost the same extent that the expected false negative rate overestimated false negative misclassifications. The false positive and false negative difference data are thus very similar, and the relatively small differences between observed and expected total misclassifications reflect the tendency for the two types of error to cancel one another.

The low expected false negative rates found for the binomial error and Bayesian models are due to the relatively low frequency of failing scores observed with the easy subtests. The high scores also produced relatively low expected false positive rates because the portions of the ability distributions associated with passing scores which were below the criterion ability were small. Low expected false positive and false negative rates led to large differences between observed and expected false positive rates and very small differences between observed and expected false negative rates. The differences in observed and expected total misclassifications are almost entirely accounted



for by the false positive differences.

The slightly higher empirical best criterion scores on the easy subtests led to lower observed false positive and higher observed false negative rates than those found for the models. When applied to the 240 round criterion ability distribution, these relatively high criterion scores also produced lower expected false positive rates than the models but the differences between the observed and expected rates were similar to the models' differences. The expected false negative rates associated with the higher empirical best criterion scores were much higher than those found for the models, leading to large observed versus expected differences. The false negative results are also reflected in high expected total misclassification rates and large differences between observed and expected total misclassifications.

The differences between observed and expected misclassifications found for the models on the easy subtests illustrate the importance of considering the relative importance of false positive and false negative errors. The models tended to underestimate false positives and overestimate false negatives. While the models were comparable with respect to false positives, the binomial error and Bayesian models predicted, accurately, very low false negatives while the proportion correct model was overly conservative. When the errors in prediction were summed to produce the error in predicting total misclassification, the values for the binomial error and Bayesian models were similar to those obtained for the false positives. In the case of the proportion correct model, however, the false positive and false negative results tended to cancel one another producing lower differences than the other two

models. Thus, in addition to considering the costs of false positive, false negative, and total misclassification errors, decision makers must consider which type or types of errors must be most accurately predicted.

The mix subtests' are more like what one would desire in a criterion-referenced test. Observed misclassification errors were modest and there was good agreement between observed and expected error rates. There was also little difference in the results for the different procedures. These results imply that the nature of the items included on a test, particularly the similarity between the test and the domain to which one would like to generalize, is one of the most critical factors in evaluating a test. The hard and easy subtests were not good representations of the overall domain of 240 rounds and none of the procedures produced clearly satisfactory results. The mix subtests were representative of the domain and all of the procedures worked well.

#### True Score Estimation

In addition to comparing the models on the basis of their characteristics as aids to decision making, their accuracy in estimating true scores was assessed. The results of this analysis are in Table B and are summarized in Figures 13 and 14.

Figure 13 shows the average, per test, sum of the individual examinees' squared discrepancies between the true scores estimated by the models on the basis of their subtest scores and their 240 round criterion true scores. The data are broken down by test difficulty and test length. An average discrepancy for each examinee of 20% of the number of rounds on any given test would be reflected as a squared

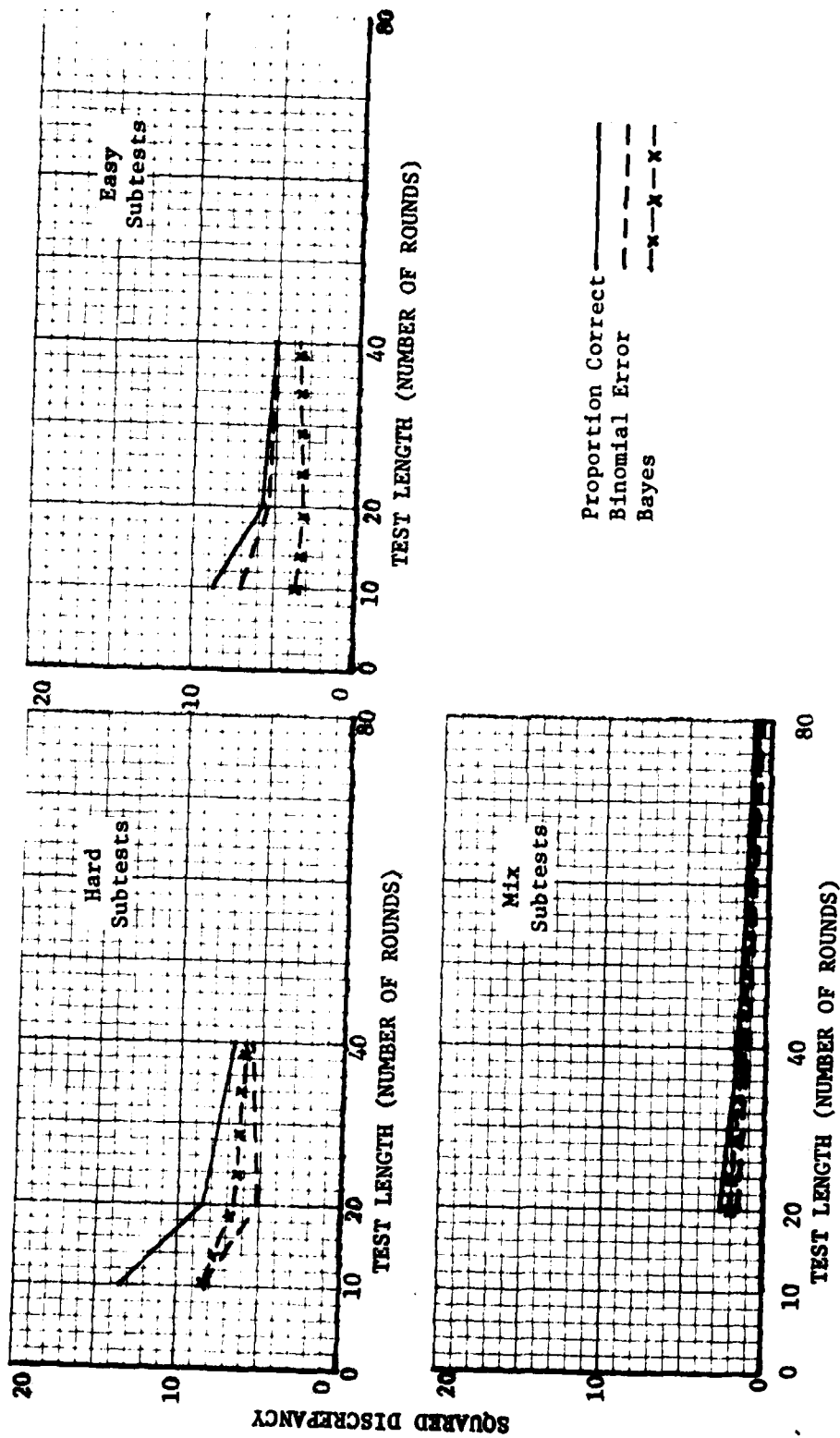


Figure 13: Average Sum of Squared Discrepancies: Subtest Estimated True Scores and 240 Round Criterion

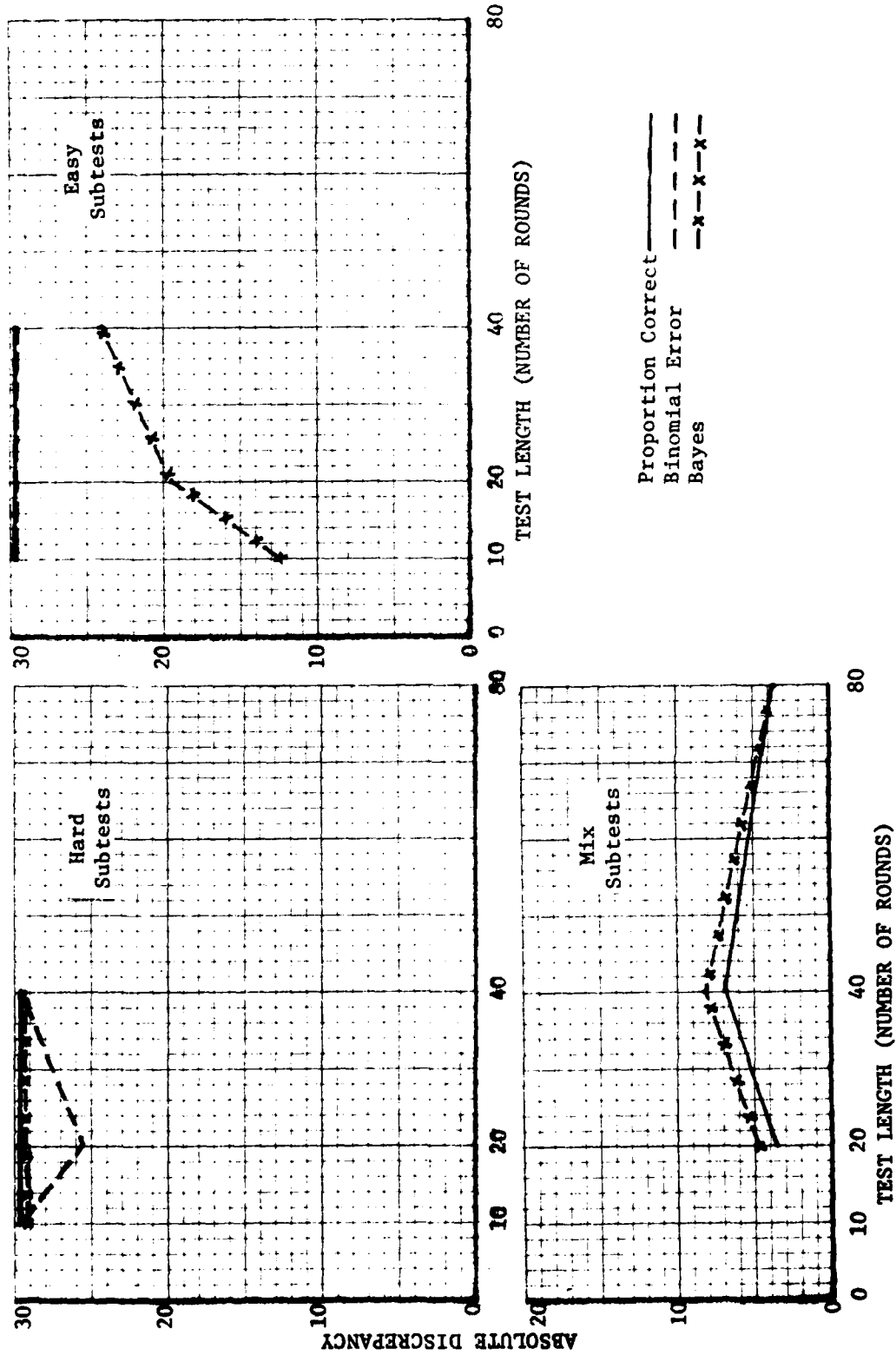


Figure 14: Average |Sum of Absolute Discrepancies|: Subtest Estimated True Scores and 240 Round Criterion

discrepancy index value of 9.48 ( $.20^2 \times 237$ ). Values for average discrepancies of 15%, 10%, and 5% are 5.33, 2.37, and .59. The majority of the average discrepancies for the hard and easy subtests thus fell in the 15% to 20% range. Most of the discrepancies for the mix subtests were in the 5% to 10% range. In all cases, the accuracy of the true score estimates improved with increasing test length. The improvement was most dramatic in going from the 10 round to the 20 round subtests after which improvement was more gradual. For all three models, the true score estimates were most accurate for the mix subtests and least accurate for the hard subtests. The squared discrepancies for the proportion correct model estimates were always higher than those for either of the other models. The binomial error and Bayesian models were very similar in their results. However, the Bayesian model tended to be slightly less accurate for the hard and mix subtests while the binomial error model was slightly less accurate for the easy subtests.

Figure 14 shows the average, per test, of the absolute values of the sum of the absolute differences between the true scores estimated by each model and the 240 round true scores. The absolute discrepancy value indicates either that the discrepancies were small in all cases or that the sum of the discrepancies for individuals whose true scores were overestimated was close to the sum of the discrepancies for individuals whose true scores were underestimated. The average absolute discrepancies tended to be relatively constant for all test lengths, with the exception of those for the Bayesian model on the easy subtests which increased as test length increased. In all cases, the

mix subtests produced less bias in estimation than the hard or easy subtests. In all cases, the models underestimated true scores for the hard subtests and overestimated true scores for the easy subtests. For the mix subtests, the average absolute discrepancy, maintaining the sign, across all test lengths was .0003 for the proportion correct model, -.0143 for the binomial error model, and -3.801 for the Bayesian model. These data indicate almost no bias for the proportion correct model, a slight tendency for the binomial error model to underestimate true scores, and a more appreciable bias towards underestimating true scores in the case of the Bayesian model. The amount of bias for the proportion correct model was nearly identical to that for the binomial error model for all types of tests and all test lengths. The Bayesian model produced results similar to the other models for the hard subtests, it tended to be less biased for the easy subtests, and it tended to be more biased than the other models for the mix subtests.

These data imply that the Bayesian model tended to produce true score estimates that were lower than those of the proportion correct or binomial error models. The differences were negligible for the hard subtests, but the tendency to produce lower true score estimates is reflected in the lower overestimation found with the easy subtests and the greater underestimation found for the mix subtests.

Comparison of the Scoring Models:  
120 Round Hard and Easy Criteria

The analyses conducted to compare the models based on their characteristics relative to the total skill domain described by the 240 round criterion test were repeated for the 120 round hard and easy subdomains. The results are described in this section. Since the

analyses based on the subdomains were intended primarily to assess the models when the subtests were close approximations to the criterion domain, only the hard subtests were compared to the 120 round hard criterion test. Similarly, the easy subtests were compared only to the 120 round easy criterion test. Table C summarizes the results for the recommended criterion scores and the misclassification rates. Table D summarizes the results of the true score estimations.

#### Criterion Score

There were no changes for these analyses in the proportion correct model's criterion scores from those used for the analyses based on the 240 round test. This is because the procedure for choosing criterion scores using the proportion correct model is not dependent on any information outside of the binomial probability tables (see Table 3). The proportion correct model's recommended criterion scores are 7 and 8 for the 10 round subtests, 14 and 15 for the 20 round subtests, and 27, 28, and 29 for the 40 round subtests.

The binomial error model's recommended criterion scores are based on the observed score distribution for each testing occasion. Since only the criterion was changed from the 240 round domain to the 120 round hard or easy subdomain, for these analyses, without disturbing the observed score distributions of the subtests, there were no changes in the criterion scores from those for the analyses based on the 240 round criterion. The binomial error model's recommended criterion scores are 8, for eleven of the 10 round hard subtests, and 7, for the remaining 10 round hard subtest, 14, for three of the 20 round hard subtests, and 15, for the other three 20 round hard subtests, and 29,

for all three 40 round hard subtests. The 10 round easy subtests' criterion scores are 4 in two cases, 5 in one case, 6 in four cases, and 7 in the remaining five cases. Criterion scores for the 20 round easy subtests are 11 in one case, 12 in three cases, and 13 in two cases. For the 40 round easy subtests, the criterion scores are 25, 26, and 27.

The criterion scores recommended by the Bayesian model did change. The prior distributions used for the analyses based on the 120 round subdomains were the distributions of scores for the hard or easy MPFQC tables included in each subdomain that were expected by the Military Police School staff. The hard prior distribution suggested that the trainees' abilities were lower than those suggested by the prior distribution based on all of the MPFQC tables. With the Bayesian model, lower prior ability estimates require higher observed scores for a "pass" decision. Therefore, the Bayesian model's criterion scores are higher for these analyses than for the 240 round criterion analyses. The 10 round hard subtests have criterion scores of 8, the 20 round hard subtests have criterion scores of 15, and the 40 round hard subtests' criterion scores are 29. The easy prior distribution suggested that the trainees' abilities were higher than those suggested by the other prior distributions. Therefore, lower criterion scores are required to confirm a "pass" decision. The 10 round easy subtests' criterion scores are 7, the 20 round easy subtests have criterion scores of 14, and the 40 round subtests' criterion scores are 28.

The empirical best criterion scores also changed for these analyses. This is because the criterion master or nonmaster status changed



with the change in criterion domain. For the 10 round hard subtests, the empirical best criterion scores are 9 in nine cases, 8 in two cases, and 7 in the remaining case. The 20 round hard subtests' empirical best criterion scores vary from 13 to 16. The empirical best criterion scores for the 40 round hard subtests are 28 in two cases and 32 in the remaining case. These empirical best criterion scores for the hard subtests reflect the distribution of masters and nonmasters according to the 120 round hard test criterion. The nonmaster group consisted of 150 persons or 63.3% of the examinees. Thus, the maximum false positive rate was .633. The master group had 87 persons or 36.7% of the group, yielding a maximum false negative rate of .367. Under these conditions the best strategy for minimizing total misclassification would be to minimize the likelihood of committing false positive errors. The relatively high empirical best criterion scores demonstrate this strategy.

The easy subtests' empirical best criterion scores clearly show the importance of the true distribution of masters and nonmasters on the choice of a criterion score. Only 5 trainees of 2.1% of the group were classified as nonmasters by the 120 round easy criterion test. Therefore, the maximum false positive rate was .021 and the maximum false negative rate was .979. Under these conditions, total misclassification can best be minimized by choosing a criterion score that minimizes the false negative rate, in other words, a low criterion score. In addition, because there were so few nonmasters and there was little variability in their scores, several criterion scores were often found to be low enough to insure very low false negative misclassification rates and equivalent total misclassification rates.

For the 10 round easy subtests, the empirical best criterion scores vary from 0 to 7, with multiple values being the rule rather than the exception. The criterion scores for the 20 round easy tests vary from 0 to 13. In three cases (Easy22, Easy25, and Easy26), multiple empirical best criterion scores are found, all of which demonstrate the strategy of minimizing false negative errors. In the other three cases, a more moderate criterion score (10 to 13) served to provide very low, equally divided false positive and false negative errors. The subtests with the more moderate criterion scores are much more useful in discriminating between masters and nonmasters than are the other subtests which were so easy that even with a criterion score of 12 hits in 20 rounds (60%) all of the nonmasters passed. The 40 round easy subtests' results repeat what was found for the 20 round easy subtests. Easy43 is exceptionally easy, its empirical best criterion scores are anything from 0 to 25, and for all of these criterion scores all nonmasters as well as all masters passed. The other two 40 round easy tests have single empirical best criterion scores of 25 in one case and 26 in the other, which led to low and more equally divided false positive and false negative misclassification rates. When there were multiple empirical best criterion scores, only the data for the highest score were included in the analyses of observed, expected, and observed versus expected misclassification rates. The highest criterion scores were chosen for subsequent analysis because they most closely approximated the models' criterion scores and therefore are more appropriate for comparisons than the more extreme scores.

Observed Misclassification Rates

The hard test observed misclassification rate data are summarized in Figure 15. Observed misclassification rates for the hard subtests are similar for the models and the empirical best procedure. False positive and false negative misclassification rates average about .100 (10% of all classifications), with the false positive rate being slightly higher. All misclassification rates tend to decrease with increasing test length. These results are in sharp contrast to the results for the hard subtests versus the 240 round criterion (Figure 8), where the models showed extremely low false positive rates and extremely high false negative rates, and where the models' results were very different from the empirical best results. The empirical best procedure produced the lowest total misclassification rates, the binomial error and Bayesian models' total misclassification rates are almost identical to each other and slightly higher than the empirical best results. The proportion correct model was slightly less effective in producing accurate classification than the other two models. The results for the FP:FN ratios are closer to 1.0 than were found for the hard subtests versus the 240 round criterion and no procedure stood out as being superior to the others in producing FP:FN ratios close to 1.0.

The easy subtest observed misclassification rate data are summarized in Figure 16. The false positive rates are very low and similar for all of the procedures. The empirical best false positive observed misclassification rate is slightly higher for the 10 round easy subtests, reflecting the tendency to allow the false positive rate to rise to its maximum value in order to minimize false negatives. The models'

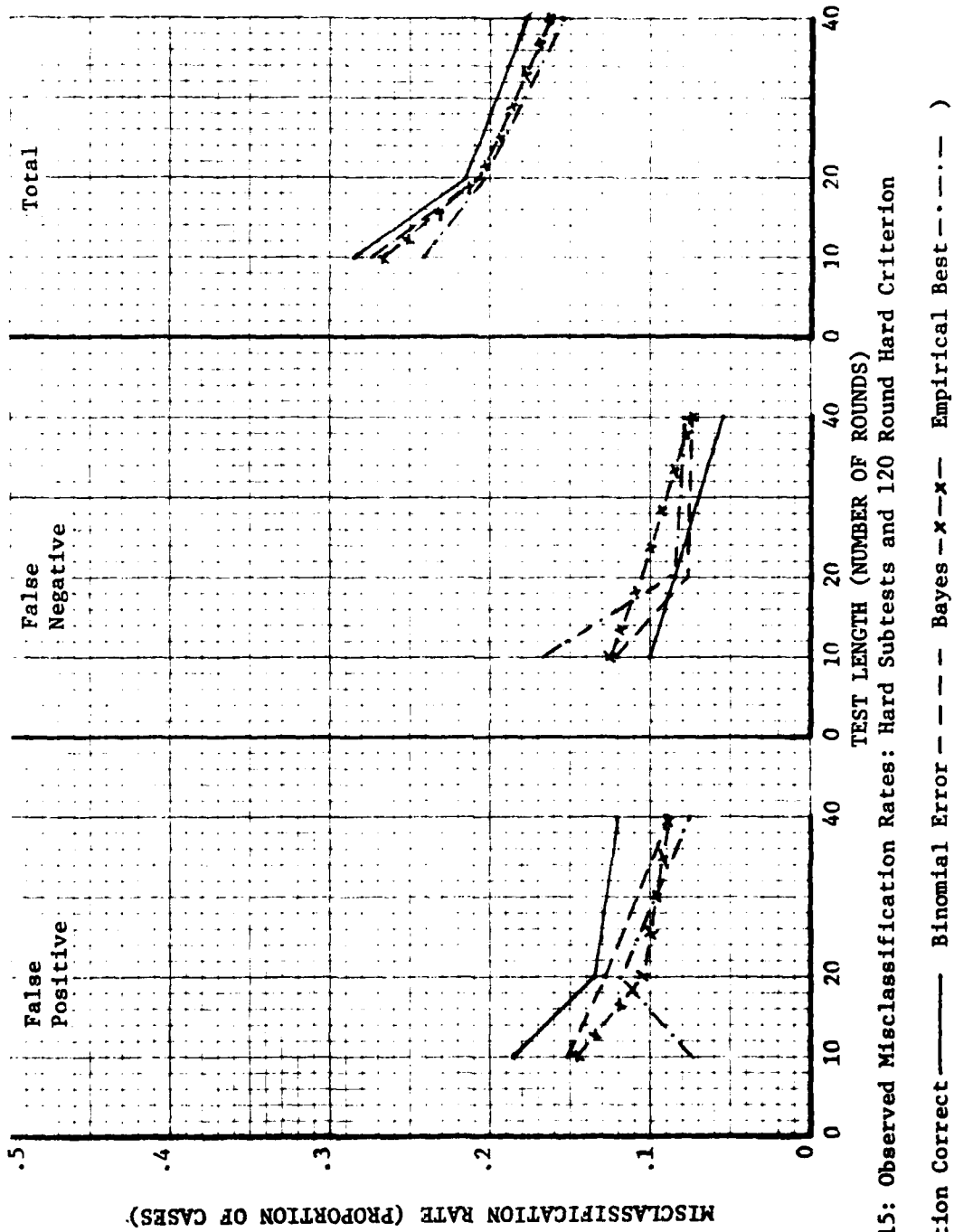


Figure 15: Observed Misclassification Rates: Hard Subtests and 120 Round Hard Criterion

(Proportion Correct ——— Binomial Error - - - - Bayes -x-x-x- Empirical Best -·-·- )

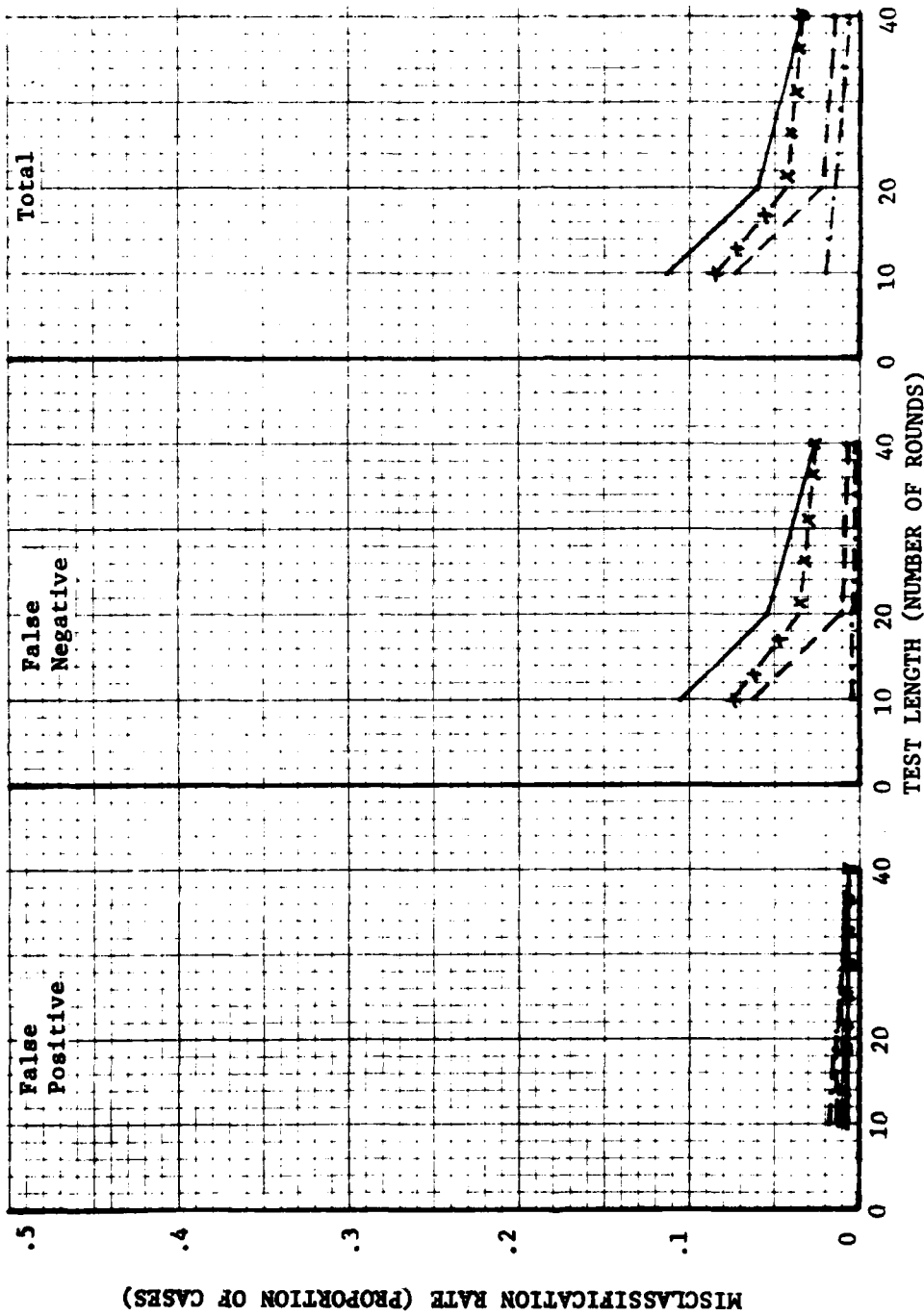


Figure 16: Observed Misclassification Rates: Easy Subtests and 120 Round Easy Criterion

(Proportion Correct ——— Binomial Error - - - Bayes - x - x - Empirical Best - · - · - )

moderately high criterion scores insured that there were very few false positive errors. The false negative misclassification rates show more variation among the procedures. As expected, the empirical best false negative rates are lowest. The binomial error model's criterion scores, which were lower than those of the other two models, produced the next lowest false negative rates and the proportion correct model, with the highest criterion scores, has the highest false negative rates. The total observed misclassification rates closely approximate the false negative results.

The easy subtests' false positive and total misclassification rates are considerably lower than the results obtained with the 240 round criterion (Figure 8). However, the large number of 120 round criterion masters produced a large pool of individuals for whom false negative misclassifications could occur. This resulted in higher false negative rates than those observed with the smaller master pool associated with the 240 round criterion.

The FP:FN ratios for the easy subtests are, in many cases, uninterpretable due to the absence of any misclassifications. Values of 0 for the false positive or false negative misclassification rates led to FP:FN ratios of 0 or undefined. Comparing the absolute differences between the false positive and false negative rates for the models and the empirical best procedure, however, shows the advantage of the empirical best strategy in producing relatively equivalent false positive and false negative misclassification rates. For the most part, the models' criterion scores produced very few false positive misclassifications but many false negative misclassifications. The empirical

best criterion scores produced low misclassification rates for both types of errors.

#### Expected Misclassification Rates

The hard subtest expected misclassification rate data are summarized in Figure 17. The easy subtest data are in Figure 18. Both sets of data indicate a high degree of similarity among the models and the empirical best procedure. Expected misclassification rates are generally low and tend to decrease with increasing test length. Differences between the procedures and between the hard and easy subtests' results are due to differences in the procedures' criterion scores and the mix of masters and nonmasters in the criterion subdomains.

For the hard subtests, the expected false positive rates are lowest for the binomial error model and highest for the proportion correct model. The expected false negative results are the reverse, the lowest false negative expected misclassification rates are found for the proportion correct model and the highest for the binomial error model. The binomial error model produced the lowest expected total misclassification rates. The proportion correct model produced the highest total expected misclassification rates for the 10 round hard subtests, but then produced results almost identical to the Bayesian model. The empirical best criterion scores were higher than those for the models on the 10 round hard subtests, producing low expected false positive rates and high expected false negative rates. The empirical best criterion scores were similar to the models' on the 20 round hard subtests, as are the expected misclassification rates. For the 40 round hard subtests, the one high empirical best criterion score is reflected

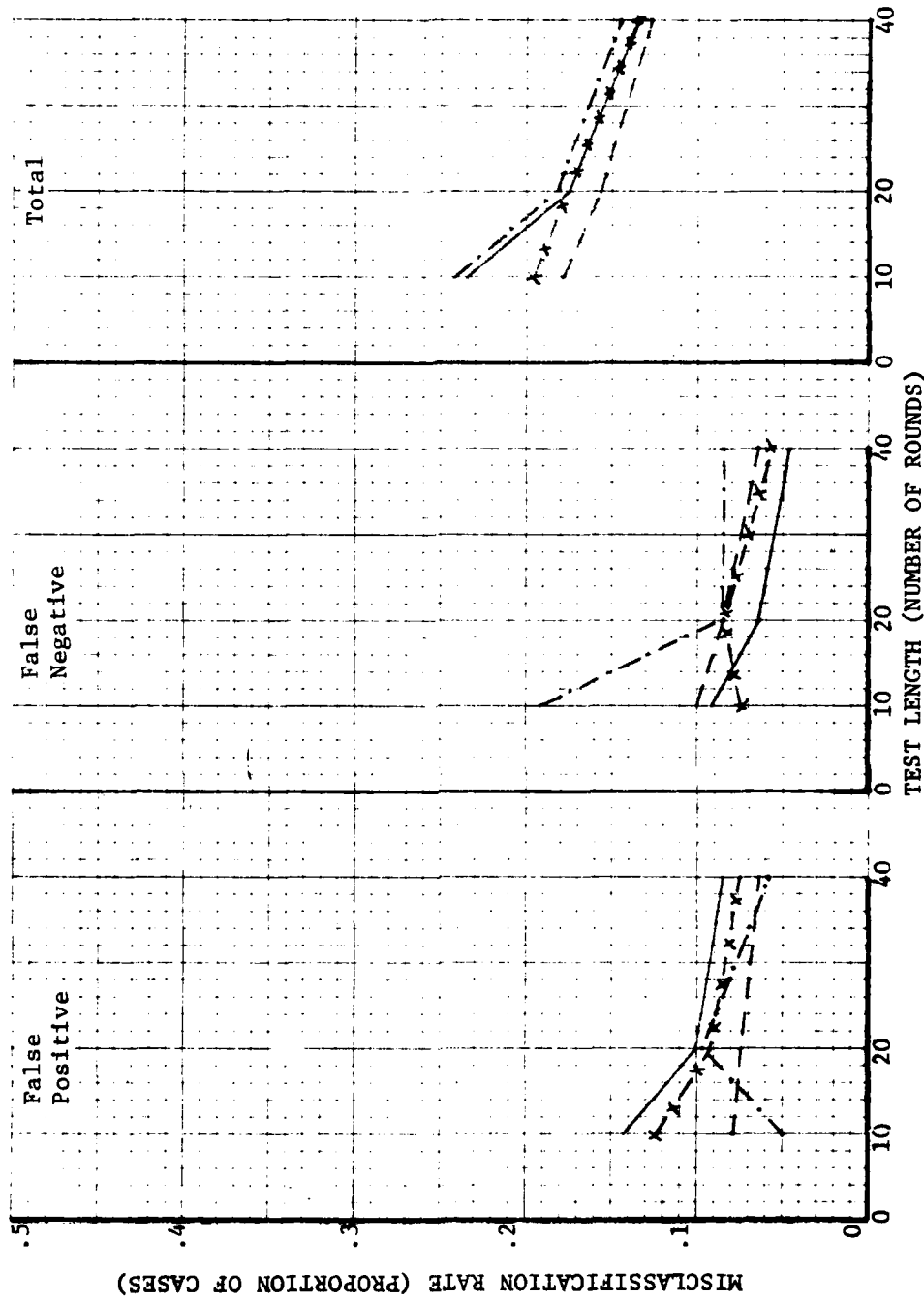


Figure 17: Expected Misclassification Rates: Hard Subtests and 120 Round Hard Criterion

(Proportion Correct ——— Binomial Error — — — Bayes — x — x — Empirical Best — · — · — )



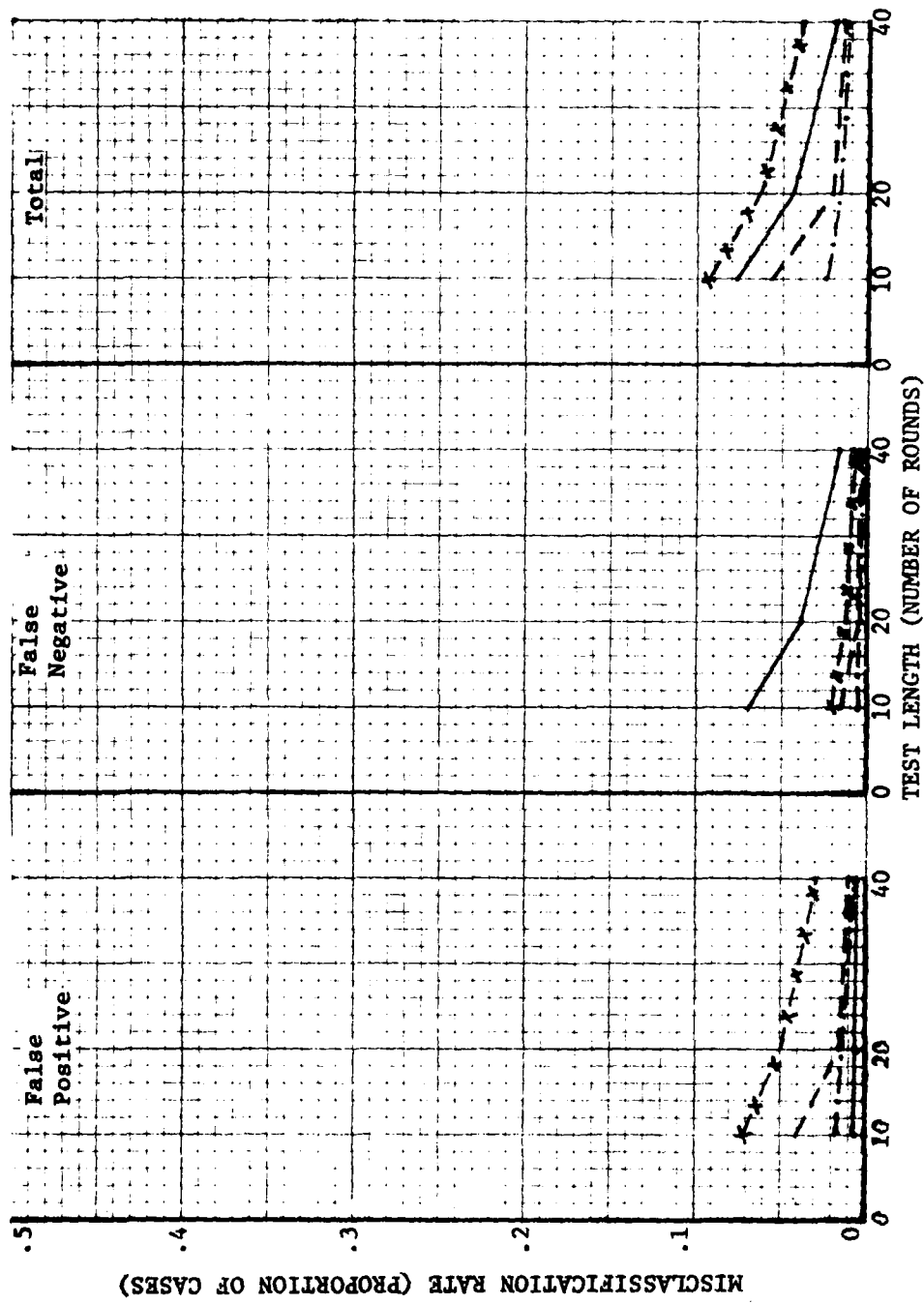


Figure 18: Expected Misclassification Rates: Easy Subtests and 120 Round Easy Criterion

(Proportion Correct ——— Binomial Error - - - Bayes - x - x - Empirical Best - . - . - )

in a relatively low expected false positive rate and a relatively high false negative rate. The total expected misclassification data for the empirical best procedure closely parallel the data for the models. This represents a canceling out of the extreme false positive and false negative values.

For the easy subtests, the proportion correct model had higher criterion scores than either of the other models. This is reflected in low false positive and high false negative expected misclassification rates relative to the results obtained for the other models. The generally lower criterion scores found for the binomial error model are reflected in relatively high expected false positive rates, but the effect disappeared in the very low expected false negative rates. The empirical best and Bayesian results are very similar for the 20 round and 40 round easy subtests. However, the empirical best expected misclassifications for the 10 round subtests are lower than those for the Bayesian model. With respect to total expected misclassification, the highest values were found for the binomial error model followed by the proportion correct and Bayesian models. The empirical best total expected misclassification rates are lower than those of the models.

The expected misclassification rates for the easy subtests are lower than those for the hard subtests. This is partly due to the difficulty of the tests and partly due to the relative proportions of masters and nonmasters as defined by the 120 round subdomains. The small proportion of nonmasters in the 120 round easy subdomain explains the very low expected false positive rates. However, this would also lead one to expect to find high expected false negative rates. This

was not the case for the binomial error and Bayesian models because the score distributions had few low scores and few examinees failed the easy subtests. Therefore, the expected false negative misclassification rates, which are proportional to the observed failing rates, were correspondingly low. The proportion correct model had higher expected false negative rates than the other models, but the examinees' abilities as defined by the 120 round easy criterion test were sufficiently high that relatively few masters would be expected to fail. Thus, the expected false negative rates for the easy subtests were lower than those for the hard subtests. The empirical best criterion scores were lower than those of the models producing the very low expected false negative rates for the easy subtests.

#### Observed versus Expected Misclassification Rates

The data describing the differences between the observed and expected misclassification rates for the hard subtests are summarized in Figure 19. Figure 20 summarizes the data for the easy subtests. The absolute values for the differences are low for all of the models and the empirical best procedure for false positives, false negatives, and total misclassifications at all test lengths and test difficulties. The differences also tend to decrease with increasing test length and tend to reflect differences of about 5% of all classifications. The hard subtests' observed false positive rates were underestimated in about 75% of the cases. For the easy subtests, the observed false positive rates were underestimated in about 30% of the cases. The hard subtests' observed false negative rates were underestimated in slightly more than 50% of the cases. For the easy subtests, the false negative rates were underestimated in about 65% of the cases. Total hard subtest misclassi-

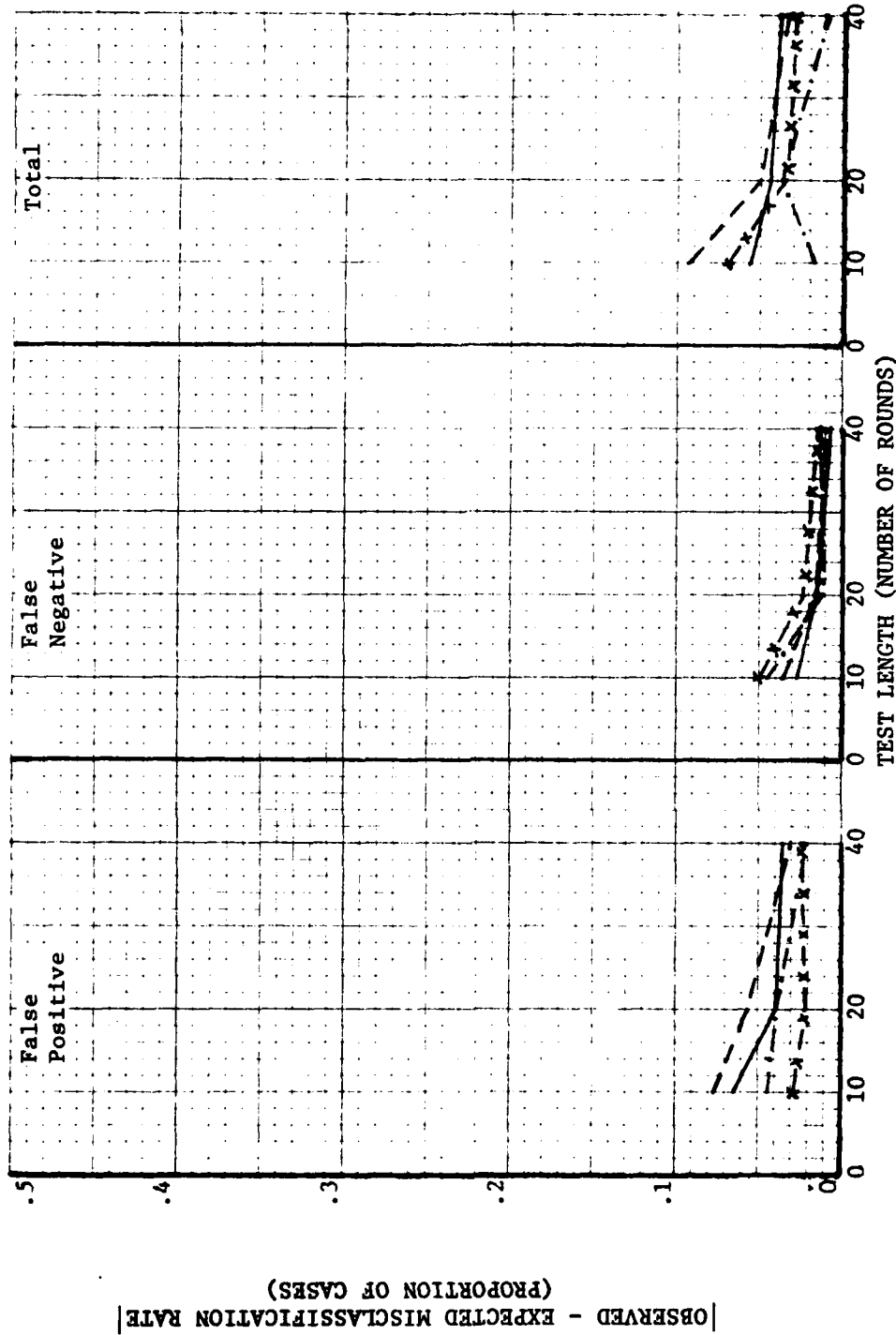


Figure 19: Absolute Values of Differences Between Observed and Expected Misclassification Rates:  
 Hard Subtests and 120 Round Hard Criterion  
 (Proportion Correct ——— Binomial Error - - - Bayes -x-x- Empirical Best - · - · - )

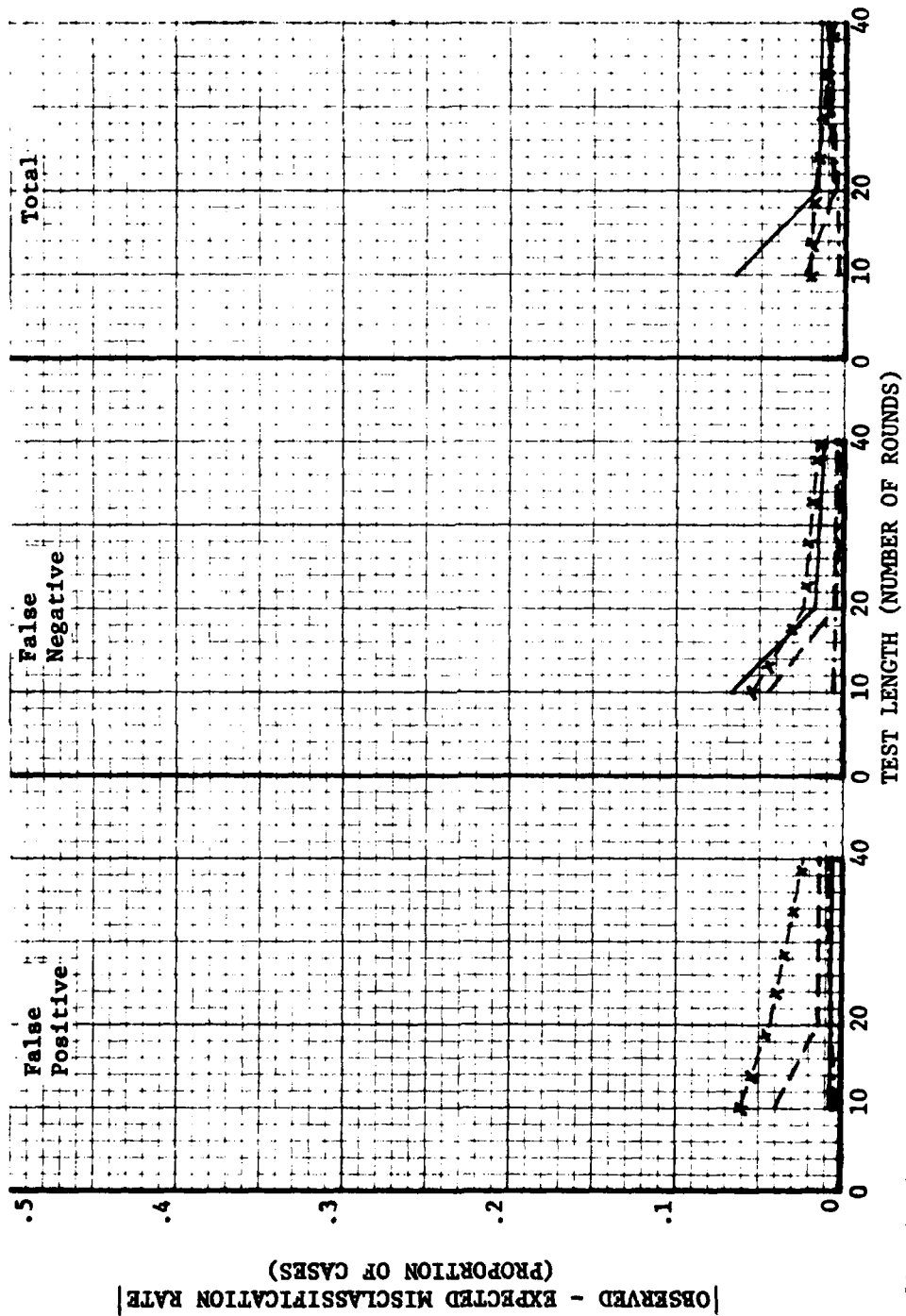


Figure 20: Absolute Values of Differences Between Observed and Expected Misclassification Rates:  
Easy Subtests and 120 Round Easy Criterion  
(Proportion Correct ——— Binomial Error — — — Bayes —x—x— Empirical Best -.-.-)

fication was underestimated in about 90% of the cases and the easy subtests' total misclassification was underestimated in about 50% of the cases.

These results are markedly different from those for the 240 round criterion (Figure 12). The differences between observed and expected misclassification rates were much lower when the subtests were matched to their appropriate subdomain, and the clear differences between the models evident in the data based on the 240 round criterion test disappeared. The models performed well, observed misclassification rates were not disturbingly high, and the models expected misclassification rates were relatively accurate representations of the observed rates.

#### True Score Estimation

Figures 21 and 22 summarize the results of the analyses to determine the accuracy of the models' true score estimates based on the subtests' scores relative to the true scores as defined by the 120 round hard (Figure 21) and easy (Figure 22) subdomains. The curves for both the hard and easy subtests are similar in shape, however, the squared discrepancies between the estimated and criterion true scores are uniformly lower in the case of the easy subtests. For both sets of data, the proportion correct model's estimated true scores were slightly less accurate than those of the other two models. The binomial error and Bayesian models were nearly identical in the accuracy of their true score estimates, with the binomial error model slightly more accurate for the easy subtests and the Bayesian model slightly more accurate for the hard subtests. In all cases, the true score estimation improved with increasing test length, the most dramatic increase in accuracy coming

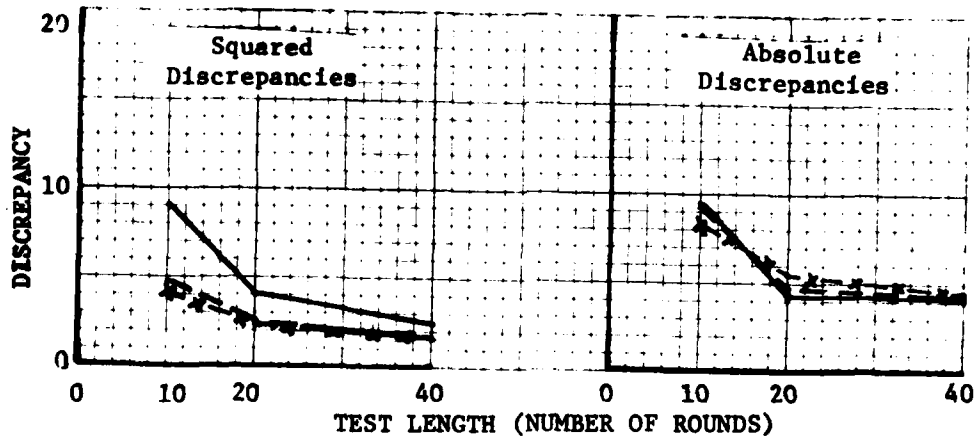


Figure 21: Average Sum of Squared Discrepancies and Average  
|Sum of Absolute Discrepancies|: Subtest Estimated  
True Scores and 120 Round Hard Criterion

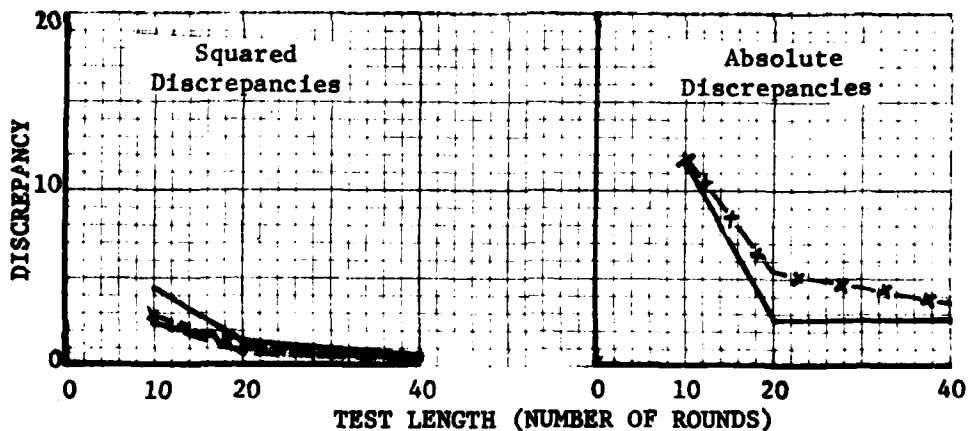


Figure 22: Average Sum of Squared Discrepancies and Average  
|Sum of Absolute Discrepancies|: Subtest Estimated  
True Scores and 120 Round Easy Criterion

(Proportion Correct ——— Binomial Error - - - - Bayes -x-x- )

between the 10 round subtests and the 20 round subtests. The proportion correct model per person per test error rates were close to 20% for the 10 round hard subtests and between 10% and 15% for the 10 round easy subtests. The 20 round subtests' results for the proportion correct model show error rates of between 10% and 15% for the hard subtests and between 5% and 10% for the easy subtests. About 10% error in estimating true scores was found for the 40 round hard subtests, and for the 40 round easy subtests the error was slightly more than 5%. The binomial error and Bayesian models' results for the hard subtests show error rates of between 10% and 15% for the 10 round subtests, error rates of about 10% for the 20 round subtests and error rates of between 5% and 10% for the 40 round subtests. The easy subtests' results show error rates of about 10% for the 10 round subtests, between 5% and 10% for the 20 round subtests, and close to 5% for the 40 round subtests. These results show considerably better estimation of true scores than was the case when the criterion true scores were based on the 240 round criterion (Figure 13).

The bias in estimating true scores, as reflected in the average absolute discrepancies, was also less when the hard or easy subtests' true score estimates were compared to the 120 round hard or easy criteria than when the 240 round criterion was used (Figure 14). The relative degree and direction of bias among the models, however, was unchanged. The proportion correct and the binomial error models were almost identical in the amount of bias in estimating true scores and the bias was very small. In contrast to the results based on the 240 round criterion, where the models consistently underestimated true scores for the hard subtests and overestimated true scores for the easy subtests, the results



based on the 120 round hard or easy criteria show little bias in either direction for the proportion correct and binomial error models. The Bayesian model's results indicate that it underestimated true scores in all cases. These results are consistent with those found with the 240 round criterion, however, the degree of bias was less when the 120 round hard or easy tests defined the criterion true scores.

## 5. DISCUSSION

This study was designed with two primary purposes in mind. The first was to investigate the characteristics of an example of an apparently well constructed criterion-referenced test. The second was to choose several statistical models that could be used to set criterion scores and estimate true scores on a criterion-referenced test, and then to compare those models on the basis of the accuracy of their implied pass or fail decisions and the accuracy of their true score estimates. In the process of carrying out the study, it became necessary to choose or develop analysis techniques which would aid in accomplishing the study's purposes. This section discusses the methods used, the results of the analyses, and the conclusions implied by the results. Some suggestions for practitioners who must deal with criterion-referenced tests are also offered.

In interpreting the results of this study, several features of the data base must be considered. These data represent performance on a well defined psychomotor task as measured by a relatively high reliability performance test. While the general principles of criterion-referenced test evaluation suggested by these data are probably generalizable, it is impossible to say whether the specific quantitative results or the relationships between the scoring models would be reproduced with different tasks or tests. An important area for continued research is to extend the methodology for empirically demonstrating the

validity of statistical models for criterion-referenced testing to other tasks and additional domains of learning.

#### Analysis of the MPFQC

The example of a criterion-referenced test which was chosen to provide the data base for this study is the Military Police Firearms Qualification Course. The MPFQC was analyzed using simple, well known techniques. The results clearly show that such techniques can provide useful information for evaluating criterion-referenced tests.

The first step in the analysis of the MPFQC was to determine the purpose of the test and to carefully define the skill domain that was to be assessed by the test. This was accomplished by discussing the test with its designers, staff members at the U.S. Army Military Police School, and by studying the items included on the test. The MPFQC was designed as a criterion-referenced performance test to certify military trainees in .45 caliber pistol marksmanship. The items on the test, referred to as tables, are shooting tasks fired from a variety of distances to the target and shooting positions. The tables were chosen to represent the kinds of problems that military police face on the job. Based on a consideration of job performance requirements, manpower needs, and other demands of the school and the Army, the MP school decided that in order to be certified, a trainee had to achieve a score of 35 hits in 50 shots.

This sort of analysis helps the evaluator understand why a test was constructed and what its purposes are. It also identifies questions. For example, on the MPFQC one question was why the maximum range was 35 meters. Another was why there were tables at 25 meters but none at

20 meters. the answers, in this case, involved practical considerations. However, in other cases, such questioning can lead to decisions to increase or decrease the number or types of items included on a test. The conclusions of this analysis were that the MPFQC appeared to be a well designed test, that the test items represented the skill domain adequately, and that the results of testing should provide valid criteria for certifying the skills of military police trainees.

The second step in the analysis of the MPFQC was to administer the test to a representative group of trainees. For this study, the test was modified to increase the number of rounds fired to a total of 240, to be fired in three independent 80 round repetitions. In general, it is best to administer any test which is being evaluated in a manner as close as possible to its actual intended use. However, the modifications imposed on the MPFQC administration were required in order to address the second purpose of this study, and it was felt that the modifications would not disturb either the properties of the test or the certification process.

Third, the trainees' test scores were analyzed. Simple descriptive statistics; means, medians, modes, variances, frequency distributions, test characteristic curves, and reliabilities were computed on the total 240 round scores, the three 80 round repetition scores and the scores obtained on each of the individual tables. The results of the analyses indicated that overall the test was moderately difficult and reliable. Taking advantage of the modification to the normal testing procedures, it was also possible to compare the results of the independent repetitions. These results indicated a slight improvement in scores over time.

The most dramatic result, however, was the unexpected finding that the tables broke down into two clearly distinct groups. The four tables shot at the longer ranges were nearly identical to one another with respect to all of their descriptive statistics, but they were very different from the four tables shot at the shorter ranges which were, again, very similar to each other. The implication is that the MPFQC is actually made up of two distinct tests, a long range hard test and a short range easy test. Further, it became clear that by choosing tables appropriately it was possible to provide three indices of marksmanship ability. If one built a test consisting of all eight tables, then a general marksmanship score could be obtained. A test made up exclusively of short range tables would provide a score on short range marksmanship skills, while a test made up exclusively of long range tables would provide a score on long range marksmanship skills. It was clear from the score distributions that a high score on the short range test did not necessarily imply that an individual would qualify on either the overall test or the long range test.

These analyses and results suggest a number of lessons for criterion-referenced test evaluators. First, simple descriptive statistics and the classical KR-21 reliability index can be meaningful. With respect to reliability, one point must be made. It was expected that the trainees would have real differences in their shooting abilities. Therefore, the test results were expected to reflect true ability differences, a requirement if classical reliability indices are to be interpreted in their usual way. In some instructional circumstances, such as when a criterion-referenced test is to be used in a mastery learning setting,

the assumption that examinees will differ with respect to their abilities to accomplish the task being tested may not be valid. In such cases, classical reliability indices will tend to have values close to zero and may behave in unpredictable ways. Epstein and Knerr (1976) discuss this effect and suggest that it may still be worthwhile to compute reliabilities as long as one is careful in interpreting the results. Simple descriptive statistics such as means and frequency distributions are interpretable regardless of the true abilities of the examinees. Their value lies primarily in flagging unanticipated results. In the case of the MPFQC, these statistics indicated that the assumption that the skill domain was homogeneous was in error, and they helped to define the components of the two subdomains. They also uncovered the slight practice effect. In other cases, such descriptive analyses can be used to identify unusual test items or to verify that instruction was uniformly effective throughout a skill domain.

Because of the design of the MPFQC administration for this study, additional analyses that might either be impossible or optional under other conditions were conducted. These included creating 20 and 40 round subtests and 120 round hard and easy criterion tests by sampling results from the pool of 240 rounds, comparing the data from the tests of differing length and difficulty, and performing an analysis of variance. The results of these additional analyses confirmed the results of the primary analyses.

The subtests were divided into groups of those containing only the hard MPFQC tables, only the easy MPFQC tables, and a mix of hard and easy MPFQC tables. The comparisons of the tests showed that test

characteristics were very similar within a difficulty type, but that they were clearly distinct across difficulty types regardless of test length. These results support the conclusions that the MPFQC represents two sub-domains and that it is reasonable to consider the three interpretations of the test scores.

The analysis of variance was used to assess the relative effects of the factors which defined the test administration procedures on the variability of the observed scores. The results showed that the majority of the observed variance in scores was due to individual differences between trainees and to differences in the tables, confirming the results of the other analyses. The analysis of variance also pointed out several important methodological considerations. The first dealt with the choice of random and fixed factors in the analysis. Many examples of the use of the analysis of variance technique treat subjects as the only random factor, with all treatment or experimental factors treated as fixed effects. For this study, that was not an acceptable design since it was desired to generalize the results beyond the specific administration constraints imposed by this study. The point is that if analysis of variance techniques are chosen as one of the methods to evaluate a criterion-referenced test, great care must be taken to insure that the design and the designation of fixed and random factors are appropriate. An additional methodological concern involves sample size. This study included 237 subjects. The large sample size contributed to a very large number of degrees of freedom in the analysis of variance error terms. Under such conditions, rejection of the statistical null hypothesis, even though experimental effects may be very small, is virtually

a certainty. Thus, one can find, as this study's results show, statistically significant F-ratios for experimentally trivial results or for factors which account for only a small proportion of total variance. The point here is that care must be taken in interpreting analysis of variance results and that subsequent analyses showing the proportion of variance accounted for by each main effect and interaction are often worth pursuing. This is particularly the case for statistically powerful experiments.

The final step in the analysis of the MPFQC, as it would be in any evaluation of a test, is to report the results and to suggest some areas that might be considered in revising the test. The recommendations, in this case, fall into two categories, interpreting the scores and revising the tables. With respect to score interpretation, the first point is that the overall score achieved on the MPFQC can be misleading. The analyses of the test showed that it is possible to achieve a high score on the easy tables, achieve only a poor to moderate score on the hard tables, and still be certified as a qualified marksman. This could result in allowing military police trainees to graduate from the school who would not necessarily perform adequately on the job. Three alternatives seem feasible. One is to raise the criterion score so that good performance is required on all tables in order to be certified. The second is to separately score the easy and hard tables to insure that adequate performance is demonstrated at all ranges. The third alternative is to have different criteria for the different tables. For example, it may be that the accuracy required at short ranges is greater than that required at long ranges. In this case, perhaps an 80% hit rate could be required for the short range tables and a 60% hit rate could be adequate



for the long range tables. With respect to revising the tables, the school should be made aware that the easiest of the easy tables, the one fired at a range of 7 meters, is so easy that nearly everyone hits the target all of the time. It therefore provides little or no information in discriminating good marksmen from poor marksmen. One possible revision could be to eliminate the 7 meter table and replace it with one that provides more useful information. A second alternative that might be considered is redistributing the number of rounds fired from each table. The recommended strategy in this case would be to decrease the number of rounds fired from short range and increase the number fired from long range. This would maintain the job relevancy of the overall test while probably increasing the power of the test to discriminate between good marksmen and poor marksmen.

#### Comparison of The Models

Setting criteria for passing criterion-referenced tests remains one of the most controversial issues in the literature (see Glass, 1978, for example). The problem has two important facets. First, one must decide what level of achievement, in an abstract sense, is necessary. In other words, if it were possible to test individuals in an ideal setting where measurement errors, time constraints, poor test items, and other disturbing factors did not exist, what levels of performance would be required? The choice of such ideal achievement levels involves both subjective judgments and consideration of what will be required of examinees who are certified competent. For example, if one was concerned with achievement in American history, the ideal achievement level might represent what a group of concerned citizens thought was necessary for good

citizenship. If the history course was part of a sequence of courses, then the ideal achievement level might also reflect the entry skills and knowledges for the next course. In a more job related setting, the ideal achievement level for an industrial assembly task might reflect what was required to insure that there was no delay on an assembly line and that quality control standards for the industry could be maintained.

In many cases, the ideal achievement level will have to be reduced for reasons completely external to testing. For example, while it might be considered important that elementary school students be able to spell all of the words introduced to them in a block of instruction, previous experience may have shown that, for many students, there is simply not enough time to learn all of the words. Under these conditions, the ideal, in the sense of no measurement error, achievement level might be that 70%, 80%, or 90% of the words needed to be spelled correctly. The point is, an ideal achievement level must be defined as the desired level of achievement assuming that there are no errors of measurement. In the case of the MPFQC, the school decided, after considering the job requirements and the practical constraints under which the training operated, that if a trainee could hit the target 70% of the time that would be an acceptable level of achievement.

The second facet of the problem is setting a criterion score on an actual test with a finite, usually small, number of test items which recognizes that errors of measurement do occur. Statistical models may help in solving this problem. The second purpose of this study was to consider how effective three statistical models were in suggesting criterion scores that led to valid pass or fail decisions and how accurately the models estimated examinee true scores.

Three models were chosen for study. They share the binomial probability distribution for describing the expected distribution of observed scores given an examinee's true ability. True abilities and estimates of examinee true scores are on a scale from 0 to 1.0, where 0 implies that the probability is 1.0 that the examinee will fail all test items and where 1.0 implies that the probability is 1.0 that the examinee will pass all test items. The models differ in the kinds of information needed in addition to the binomial probability distribution to compute a recommended criterion score and to compute estimated true scores.

The first model is being referred to as the proportion correct model. The additional information required for setting a criterion score is the subjective judgment of an evaluator, teacher, test designer or other informed person or group of persons. Examinee true score estimates follow directly from the observed performance. The model simply states that the expected distribution of observed scores given a true ability, which is defined as the probability of answering any test item correctly, is the binomial distribution. A criterion score is chosen by considering the relative probabilities that examinees above and below the ideal achievement level will obtain scores at least equal to candidate criteria. The desired case is to find a criterion score for which the probabilities are high that examinees at or above the ideal achievement level will obtain at least that score, and for which the corresponding probabilities for examinees below the ideal achievement level are small. The probabilities for examinees below ideal achievement are interpreted as the probabilities of committing false positive decision errors. One minus the probabilities for examinees at or above ideal achievement are

interpreted as the probabilities of committing false negative decision errors. According to the proportion correct model, each examinee's true score estimate equals the proportion of items on a test answered correctly, or, for the MPFQC, the proportion of shots that hit the target.

The second model is the binomial error model. This model begins at the same point as the proportion correct model. However, it differs in the procedures for recommending criterion scores and estimating examinee true abilities. The model shows that it is possible, on mathematical grounds, to prove that the observed scores for a group of examinees are linearly related to the true scores for those examinees, and that the true scores are described by a beta distribution. The proof holds if the observed score distribution for given ability is binomial and if the observed score distribution across the group consisting of individuals with a variety of abilities is negative hypergeometric. Since many observed score distributions can be shown to fit one of the members of the family of negative hypergeometric distributions, the model can be applied in many cases. In practice, one simply computes the necessary parameters and applies the linear equation relating observed to true scores to each observed score. The output of this procedure is a set of true score estimates corresponding to the observed scores. The recommended criterion score for any given test administration is the lowest observed score corresponding to a true score estimate at or above the ideal achievement level. The probabilities of false positive and false negative decision errors are related to the distributions of errors of estimation for each score. Each estimated true score can be interpreted as the mean of a beta distribution of true scores. For each failing

score, the portion of this distribution above the ideal achievement level represents the probability that individuals with abilities above the ideal achievement level were incorrectly failed, false negative decision errors. The probabilities of false positive decision errors are equal to the portions of the distributions for passing scores below the ideal achievement level.

The third model is the Bayesian beta-binomial model. This model assumes that a binomial distribution describes observed scores given true abilities and that true abilities are distributed according to a beta distribution. These are the same assumptions as those that underlie the binomial error model. However, rather than relating the true score distribution to the observed score distribution directly, as the binomial error model does, the true score distribution that is believed to be the case is specified before data is collected and is incorporated into the decision making process. This distribution is called a prior distribution. The mathematics of the Bayesian model takes the beliefs expressed by the prior distribution, modifies them on the basis of observed scores, and produces posterior distributions which describe the distributions of abilities corresponding to each observed score. The choice of a criterion score and the true score estimates are based on these posterior distributions. In practice, one finds the lowest observed score whose posterior distribution implies that the probability is at least .5, or some other value if the relative costs of false positive and false negative errors differ, that an individual's ability equals or exceeds the ideal achievement level, and chooses that score as the criterion score. The means of the posterior distributions define the true

score estimates corresponding to each observed score. For this study, prior distributions were obtained by asking military police school marksmanship instructors what they thought the observed score distributions would look like based on their experience.

The results of these analyses showed a remarkable degree of similarity among the models. Recommended criterion scores were, for the most part, in the 70% to 80% hit range. For some of the easy subtests, the binomial error model recommended criterion scores lower than those of the models, reflecting the apparently high ability levels demonstrated by the high scores obtained on the easy subtests. Differences were also found in the criterion scores recommended by the Bayesian model as a function of the prior distribution used. When the prior distribution was based on all eight MPFQC tables, the criterion scores were 70% to 75% hits. When the prior distribution was based on the four hard MPFQC tables, the criterion scores rose to 75% to 80% hits, reflecting the instructors' beliefs that trainees would not appear as proficient on the hard tables as they would overall. The instructors' beliefs that the trainees would appear to be more proficient if only data from the four easy MPFQC tables were considered, were reflected in the prior distribution based on the easy tables. The criterion scores, in that case, were 70% hits. Despite the small differences among the models' criterion scores, the overall impression of the results of these analyses, is that it doesn't make much difference which model is used.

Regardless of which model is chosen, it is important to have some feel for how good the decisions based on the criterion scores are. This was explored by comparing the pass/fail decisions on the 10, 20, 40, and

80 round subtests with the pass/fail decisions based on a criterion of 70% hits on all 240 rounds and on criteria of 70% hits on the hard and easy 120 round tests. In addition, a baseline, empirical best, criterion score was defined as that score which produced the least amount of misclassification error.

The comparison analyses were broken down by test length and test difficulty. This was done to illustrate the effect of test length on decision making accuracy and to determine whether it was important to match the subtests and the longer criterion tests with respect to the difficulties of the MPFQC tables represented. The results of the comparisons were clear and consistent. When the difficulty of the subtests did not match that of the criterion test, decision making error rates were high, the false positive and false negative rates were very different from one another, and the empirical best criterion scores were usually not the same as the models' criterion scores. However, if subtests and criterion tests were matched, the results were just the opposite. Decision making error rates were low and equally divided between false positive and false negative errors, and the empirical best criterion scores were often recommended by at least one of the models.

Mismatches between subtests and criteria were the case when the hard and easy subtests were compared to the 240 round criterion. Since relatively low scores were obtained on the hard subtests, relatively few false positive errors were observed, but false negative and total misclassification was high. Approximately 3% of all classifications were false positives, 35% of all classifications were false negatives, and 38% of the classifications represented either a false positive or a

false negative error. Scores on the easy subtests were relatively high. This was reflected in high false positive rates, about 20% of all classifications, and low false negative rates, about 3% of all classifications.

In order to compensate for the mismatches in difficulty, the empirical best criterion scores were lower than those of the models in the case of the hard subtests and higher than those of the models in the case of the easy subtests. The misclassification error rates obtained with the empirical best criterion scores were typically more evenly divided between false positive and false negative errors than for the models. The results for the hard and easy subtests were also comparable. False positive misclassifications occurred for about 13% of all classifications, about 5% of all classifications were false negatives, and about 18% of all classifications represented decision errors.

Examples of well matched subtests and criteria were the mix subtests compared to the 240 round criterion, the hard subtests compared to the 120 round hard criterion, and the easy subtests compared to the 120 round easy criterion. The results for these comparisons generally showed minimal differences between the models and the empirical best procedure, misclassification error rates were low, and false positive and false negative errors were relatively equal. About 9% of all classifications represented false positive errors and 10% of all classifications were false negative errors, yielding a 19% overall error rate when the mix subtest classifications were compared to the 240 round criterion classifications. The results obtained when the hard subtests were compared to the 120 round hard criterion showed about 14% false positive errors, 10% false negative errors, and 24% error overall. For the easy subtests compared to the 120 round easy criterion, about 1% of all



classifications were false positive errors, about 6% of all classifications were false negatives, and about 7% of all classifications represented decision errors.

These results imply that regardless of the choice of statistical model, a relatively large proportion of decisions represent incorrect master/nonmaster classifications when a criterion-referenced test does not match the skill domain, but that relatively accurate classification can be obtained if the match is good. Unfortunately, these results do not tell the complete story because they do not fully incorporate the relative proportions of masters and nonmasters in the examinee group.

According to the 240 round criterion, about 26% of the examinees were nonmasters and about 74% were masters. Therefore, the 3% false positive rate observed with the hard subtests also implies that about 12% of the nonmasters were misclassified as masters. In the case of the easy subtests, about 77% of the nonmasters were misclassified as masters. The false negative rates, interpreted in this way, imply that about 46% of the masters were incorrectly failed on the hard subtests, but that only 3% of the masters were misclassified on the easy subtests. The mix subtests' results imply that about 33% of the nonmasters were misclassified and that about 14% of the masters were misclassified.

According to the 120 round hard test criterion, about 63% of the group consisted of nonmasters and 37% were masters. The implications are that about 23% of the nonmasters were misclassified and about 26% of the masters were misclassified when the decisions based on the hard subtests are compared to the 120 round hard test criterion classifications. In the case of the 120 round easy test criterion, about 2% of

the examinees were nonmasters and the other 98% were masters. Therefore, about 38% of the nonmasters incorrectly passed the easy subtests but only 6% of the masters incorrectly failed. These results also show that a good match between criterion-referenced tests and the criterion skill domain produces more accurate classification than a poor match, but the results are not as dramatic as those which consider only the relative proportions of all classifications which are errors.

The primary reasons for using criterion-referenced tests are to provide results which are interpretable in terms of what examinees can and cannot do and to provide data for valid classification decisions. The skills required by the MPFQC fulfill the first objective in that there is no question that the test can be interpreted in terms of examinee marksmanship ability. However, the results of the classification analyses suggest that decision error is likely to be a source of problems, at least for tests with a reasonable number of items. Until procedures which are more effective than the statistical models considered in this study are developed, it appears that the best solution is to be aware of the factors which influence the accuracy of classification decisions and to interpret test results with caution.

The two most important factors identified in this study are the apparent difficulty of a criterion-referenced test relative to the difficulty of the skill being measured and the proportions of masters and nonmasters in the examinee group. Users of criterion-referenced tests should consider what they expect test results to look like so that if unanticipated results do occur, they can be interpreted. For example, imagine that the Military Police School decided to revise the MPFQC to eliminate the four easy tables. Past experience with the original test

probably suggested that most examinees were able to achieve qualifying scores, but that extremely good scores were rare. With these expectations, the results of the hard subtests can be interpreted. The relatively low average scores and large number of fail decisions should imply that many of the decisions are false negative decisions, but that very few nonmasters are passing. Given this interpretation, it may be desirable to lower the criterion score, with the understanding that such a move would be likely to increase the number of false positive misclassifications in return for decreasing the number of false negative misclassifications. Predicting test results in advance can also be useful when new tests are being field tested. Consider the use of a criterion-referenced test for pre- and post-instruction testing. At the time of the pre-test, the skill domain would be expected to represent difficult tasks for the examinees and most of them would be expected to be nonmasters. This situation is approximated by the MPFQC 120 round hard subdomain. The results would be expected to show a low average score with relatively few persons passing. At the time of the post-test, most examinees should have mastered the skills and should find the test easy, a situation approximated by the MPFQC 120 round easy subdomain. In this case, the average score should be high and few examinees should fail. If the test results are very different from the expectations, then the validity of the assumptions concerning the test items or the abilities of the examinees must be considered. A pre-test that appears to be too easy may mean that the test items are poorly constructed and contain hints or that the more difficult skills in the domain are not included or not represented in sufficient numbers by the test items. In either case,

the test should be revised. If the problem does not appear to lie in the test items, the implication is that many of the examinees have already mastered the material. Unexpected results on the post-test may imply problems in the test items, or, particularly if the pre-test results were reasonable, they probably imply that the instruction was not as good as that desired.

While the solution to misclassification problems may not lie in the statistical models included in this study, they can support the essentially intuitive analysis of expected and observed results discussed above. This was investigated by comparing the observed misclassification errors with the amount of misclassification predicted by the statistical properties of the models. The results of these analyses confirmed those of the previous analyses. The differences between models are relatively small, and one is much better off when tests are matched to their criterion domains.

When decisions based on the hard subtests were compared to the 240 round test decisions, the models predicted, on the average, about two and a half times as many false positive errors and about three times too few false negative errors as were observed. The results for the easy subtests compared to the 240 round test showed that the models' predictions averaged about three times too few false positive errors and about two and a half times too many false negative errors. In other words, when the tests did not match the skill domain, the magnitudes of the error rates were unpredictable. The directions of over and under estimation were, however, what would be expected on intuitive grounds. Fewer false positives and more false negatives were observed than were expected in

the case of the hard subtests, and more false positives and fewer false negatives were observed in the case of the easy subtests.

When the tests better matched their criteria, the theoretical results were much more similar to the observations. The models' average predicted false positive misclassification rate was slightly less than what was observed when the mix subtests' decisions were compared to the 240 round criterion decisions, and approximately one and a third times too few false negative errors were predicted. The models predicted, on the average, one and a third times too few false positive errors and one and a quarter times too few false negative errors as were observed in the case of the hard subtests compared to the 120 round hard criterion. The very low error rates observed with the easy subtests compared to the 120 round easy criterion were not predicted as well as those for the other examples of a close match between a test and its criterion. About three times as many false positive errors were predicted as were observed and about two times too few false negative errors were predicted.

The final criterion used to compare the statistical models was how closely the true scores estimated by the models based on the subtest scores approximated the 240 round and 120 round criterion true scores. Since true score, as defined for this study, is directly interpretable in terms of the probability that an examinee can display the skill being measured, accurate true score estimation is highly desirable in a statistical model designed to support criterion-referenced testing. The results paralleled those of the other analyses. There is little difference between the models, but there is considerable difference in the results obtained for subtests that do and do not match their criterion domains.

When the true scores estimated on the basis of the hard and easy subtests were compared to the 240 round criterion true scores, errors in the range of 15% to 20% were found. When the hard subtest true score estimates were compared to the 120 round hard criterion true scores, the error rates fell to between 10% and 15%. The results obtained when the easy subtests' estimated true scores were compared to the 120 round easy criterion true scores showed a drop in the error rate to between 5% and 10%. The other example of a close approximation between tests and criterion, the mix subtests' results compared to the 240 round results, also showed error rates in the 5% to 10% range.

Bias in predicting true scores was also much less when tests were well matched to criteria. The 240 round criterion true scores were grossly underestimated when the models were applied to the hard subtests' results and overestimated with the easy subtests' results. There was very little bias, however, in either direction when the true scores estimated on the basis of the mix subtests' results were compared to the 240 round criterion true scores, when the hard subtests' true score estimates were compared to the 120 round hard criterion true scores, or when the easy subtests' true score estimates were compared to the 120 round easy criterion true scores.

## 6. SUMMARY AND CONCLUSIONS

A criterion-referenced performance test of pistol marksmanship was evaluated on logical and empirical grounds. The test scores, obtained by military police trainees, were then used as a data base for comparing three statistical models, the proportion correct model, the binomial error model, and the Bayesian beta-binomial model, with respect to their relative effectiveness as aids for making pass and fail decisions and their relative accuracy in estimating examinee measurement error free true scores. The results consistently led to the same conclusions. There are few practical differences between the models in terms of the amount of decision making error that is observed, the predictability of the magnitude or direction of the decision error, or in the accuracy of true score estimates based on observed test scores. The most important consideration in evaluating criterion-referenced tests and in keeping the amount of decision error to a minimum is how closely matched the test items or tasks are to the skill domain they are intended to represent.

Evaluations of criterion-referenced tests should include analyses intended to describe the skill domain, the rationale behind the choice of test items or tasks, the purpose of the test, and the reason for the level of skill chosen to represent adequate mastery of the domain. Criterion-referenced tests which do not appear to adequately represent the skill domain or which do not require sufficient performance to meet the purpose of the test should be revised. Pilot test data can be

analyzed using well known descriptive statistics or inferential techniques such as means, variances, frequency distributions, KR-21 reliabilities, and the analysis of variance, to empirically confirm or indicate errors in the interpretation of a logical analysis of a criterion-referenced test. In the case of the test evaluated as part of this study, for example, the logical analysis indicated that the test fulfilled the requirements for a well designed criterion-referenced performance test. The empirical analysis, however, made it clear that an assumption that the domain represented a unitary skill was questionable. In fact, the empirical data indicated a two-dimensional domain and suggested that test scores could be interpreted in terms of the overall domain or independently for each of the two subdomains.

The comparisons of the statistical models indicated relatively few differences between the models, and no evidence was found which would indicate that one model should be considered either superior or inferior to the others. The comparison data did, however, clearly demonstrate the importance of a close match between test items and the domain to which results are to be generalized. When test items did not match the skill domain, the risk of incorrect classification decisions was high, the magnitude of the decision errors was not accurately predicted by statistical considerations, and the true abilities of examinees were poorly estimated by the models. When the items more closely approximated the domain, the amount of classification error decreased, it was more predictable, and true abilities were more accurately estimated. The comparison data also illustrated the effect of the relative proportions of masters and nonmasters in the examinee group on the interpre-



tation of misclassification error rates. For example, if the group consists primarily of masters, a very low percentage of the classifications are likely to represent false positive errors. The low false positive error rate may, however, obscure the fact that all or nearly all of the nonmasters in the group are misclassified. Thus, decision makers must consider the relative mix of the abilities of the examinees in interpreting test results.

Decision errors will probably always be a problem when criterion-referenced tests are administered. The results of this study suggest that the most important action that can be taken to keep the magnitude of decision error to a reasonable level is to insure that the test items adequately represent the skill domain they are intended to measure. If the match between the test items and the domain is good, then the statistical models considered in this study, along with subjective estimates of the proportions of masters and nonmasters in the examinee group can be used to estimate the types, amounts, and impact of misclassification error on decision making. As far as what the most reasonable practical solution to the problem of setting criterion scores and making pass or fail judgments is concerned, Dawes and Corrigan in their 1974 paper on the use of linear models in decision making perhaps said it best, "The whole trick is to decide what variables to look at and then to know how to add" (p.105).

## APPENDIX

Tables of Criterion Scores, Observed,  
Expected, and Observed versus Expected  
Misclassification Rates, and Squared  
and Absolute Discrepancies Between  
Estimated and Criterion True Scores

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
TABLE 11	PROP CORRECT	7	.072	.266	.338	.122	.093	.215	-.050	.173	.123
	PROP CORRECT	8	.034	.329	.363	.062	.218	.280	-.128	.111	.083
	BINOMIAL ERROR	8	.034	.329	.363	.080	.102	.182	-.046	.227	.181
	BAYES	7	.072	.266	.338	.121	.056	.177	-.049	.210	.161
	BAYES	8	.034	.329	.363	.069	.105	.174	-.035	.224	.189
TABLE 12	EMPIRICAL BEST	5	.143	.076	.219	.225	.008	.233	-.082	.068	-.014
	PROP CORRECT	7	.038	.329	.367	.122	.093	.215	-.084	.236	.152
	PROP CORRECT	8	.030	.435	.465	.062	.218	.280	-.032	.217	.185
	BINOMIAL ERROR	8	.030	.435	.465	.084	.103	.187	-.054	.332	.278
	BAYES	7	.038	.329	.367	.117	.070	.187	-.079	.239	.180
TABLE 13	BAYES	8	.030	.435	.465	.058	.125	.183	-.028	.310	.282
	EMPIRICAL BEST	4	.173	.042	.215	.247	.001	.248	-.074	.041	-.033
	PROP CORRECT	7	.034	.371	.405	.122	.093	.215	-.088	.278	.190
	PROP CORRECT	8	.013	.515	.528	.062	.218	.280	-.049	.297	.248
	BINOMIAL ERROR	8	.013	.515	.528	.073	.090	.163	-.060	.425	.365
TABLE 14	BAYES	7	.034	.371	.405	.117	.070	.187	-.079	.259	.180
	BAYES	8	.013	.515	.528	.044	.133	.177	-.031	.382	.351
	EMPIRICAL BEST	4	.135	.084	.219	.247	.001	.248	-.112	.083	-.029
	PROP CORRECT	7	.051	.304	.355	.122	.093	.215	-.071	.211	.140
	PROP CORRECT	8	.034	.392	.426	.062	.218	.280	-.028	.174	.146
TABLE 15	BINOMIAL ERROR	8	.034	.392	.426	.073	.113	.186	-.039	.279	.240
	BAYES	7	.051	.304	.355	.112	.071	.183	-.061	.233	.172
	BAYES	8	.034	.392	.426	.057	.122	.179	-.023	.270	.247
	EMPIRICAL BEST	5	.135	.063	.198	.225	.003	.233	-.090	.055	-.035

Table A: Recommended Criterion Scores and Observed, Expected, and Observed versus Expected Misclassification Rates: 10 Round Hard Subtests and 240 Round Criterion

[illegible]

### Table A (cont): 10 Round Hard Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE	
			FP	FN	TOT	FP	FN	TOT	FP	FN
TABLE31	PROP CORRECT	7	.076	.249	.325	.122	.093	.215	-.046	.156
	PROP CORRECT	8	.046	.371	.417	.062	.218	.280	-.016	.153
	BINOMIAL ERROR	8	.046	.371	.417	.081	.110	.191	-.035	.261
	BAYES	7	.076	.249	.325	.149	.045	.194	-.073	.204
	BAYES	8	.046	.371	.417	.070	.118	.188	-.024	.253
TABLE32	EMPIRICAL BEST	5	.143	.097	.240	.225	.008	.233	-.082	.089
	PROP CORRECT	7	.093	.245	.338	.122	.093	.215	-.029	.152
	PROP CORRECT	8	.072	.342	.414	.062	.218	.280	.010	.124
	BINOMIAL ERROR	8	.072	.342	.414	.077	.113	.190	-.005	.229
	BAYES	7	.093	.245	.338	.134	.052	.186	-.041	.193
TABLE33	BAYES	8	.072	.342	.414	.073	.109	.182	-.001	.233
	EMPIRICAL BEST	4	.207	.046	.253	.247	.001	.248	-.040	.045
	PROP CORRECT	7	.063	.342	.405	.122	.093	.215	-.059	.249
	PROP CORRECT	8	.038	.430	.468	.062	.218	.280	-.024	.212
	BINOMIAL ERROR	8	.038	.430	.468	.073	.094	.167	-.035	.336
TABLE34	BAYES	7	.063	.342	.405	.114	.058	.172	-.051	.284
	BAYES	8	.038	.430	.468	.055	.113	.168	-.017	.317
	EMPIRICAL BEST	3	.203	.017	.220	.255	0	.255	-.052	.017
	PROP CORRECT	7	.105	.190	.295	.122	.093	.215	-.017	.097
	PROP CORRECT	8	.072	.283	.355	.062	.218	.280	.010	.065
TABLE34	BINOMIAL ERROR	7	.105	.190	.295	.134	.045	.179	-.029	.145
	BAYES	7	.105	.190	.295	.148	.044	.192	-.043	.146
	BAYES	8	.072	.283	.355	.082	.105	.187	-.010	.178
	EMPIRICAL BEST	4	.211	.030	.241	.247	.001	.248	-.036	.029
										-.007

Table A (cont): 10 Round Hard Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
TABLE15	PROP CORRECT	7	.152	.021	.173	.122	.093	.215	.030	-.072	-.042
	PROP CORRECT	8	.131	.076	.207	.062	.218	.280	.069	-.142	-.073
	BINOMIAL ERROR	6	.198	.004	.202	.086	.007	.093	.112	-.003	.109
	BAYES	7	.152	.021	.173	.145	.020	.165	.007	.001	.008
	BAYES	8	.138	.076	.207	.107	.057	.164	.024	.019	.043
TABLE16	EMPIRICAL BEST	7	.152	.021	.173	.122	.093	.215	.030	-.072	-.042
	PROP CORRECT	7	.139	.076	.215	.122	.093	.215	.017	-.017	0
	PROP CORRECT	8	.110	.139	.249	.062	.218	.280	.048	-.079	-.031
	BINOMIAL ERROR	7	.139	.076	.215	.065	.060	.125	.074	.016	.090
	BAYES	7	.139	.076	.215	.125	.035	.160	.014	.041	.055
TABLE17	BAYES	8	.110	.139	.249	.078	.080	.158	.032	.059	.091
	EMPIRICAL BEST	6	.194	.021	.215	.182	.031	.213	.012	-.010	.002
	EMPIRICAL BEST	7	.139	.076	.215	.122	.093	.215	.017	-.017	0
	PROP CORRECT	7	.203	.008	.211	.122	.093	.215	.081	-.085	-.004
	PROP CORRECT	8	.177	.038	.215	.062	.218	.280	.115	-.180	-.065
TABLE18	BINOMIAL ERROR	6	.228	.004	.232	.032	.006	.038	.196	-.002	.194
	BAYES	7	.203	.008	.211	.102	.011	.113	.101	-.003	.098
	BAYES	8	.177	.038	.215	.074	.037	.111	.103	.001	.104
	EMPIRICAL BEST	7	.203	.008	.211	.122	.093	.215	.081	-.085	-.004
	PROP CORRECT	7	.253	.004	.257	.122	.093	.215	.131	-.089	.042
TABLE19	PROP CORRECT	8	.236	.008	.244	.062	.218	.280	.174	-.210	-.036
	BINOMIAL ERROR	5	.257	0	.257	.002	0	.002	.255	0	.255
	BAYES	7	.253	.004	.257	.054	.001	.055	.199	.003	.202
	BAYES	8	.236	.008	.244	.043	.011	.054	.193	-.003	.190
	EMPIRICAL BEST	8	.236	.008	.244	.062	.218	.280	.174	-.210	-.036

Table A (cont): 10 Round Easy Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
TABLE25	PROP CORRECT	7	.156	.034	.190	.122	.093	.215	.034	-.059	-.025
	PROP CORRECT	8	.076	.072	.148	.062	.218	.280	.014	-.146	-.132
	BINOMIAL ERROR	7	.156	.034	.190	.063	.037	.100	.093	-.003	-.090
	BAYES	7	.156	.034	.190	.139	.021	.160	.017	.013	.030
	BAYES	8	.076	.072	.148	.078	.078	.156	-.002	-.006	-.008
TABLE26	EMPIRICAL BEST	8	.076	.072	.148	.062	.218	.280	.014	-.146	-.132
	PROP CORRECT	7	.122	.034	.156	.122	.093	.215	0	-.059	-.059
	PROP CORRECT	8	.093	.118	.211	.062	.218	.280	.031	-.100	-.069
	BINOMIAL ERROR	7	.122	.034	.156	.080	.028	.108	.042	.006	.048
	BAYES	7	.122	.034	.156	.132	.024	.156	-.010	.010	0
TABLE27	BAYES	8	.093	.118	.211	.073	.079	.152	.020	.039	.059
	EMPIRICAL BEST	7	.122	.034	.156	.122	.093	.215	0	-.059	-.059
	PROP CORRECT	7	.219	.034	.253	.122	.093	.215	.097	-.059	.038
	PROP CORRECT	8	.181	.055	.236	.062	.218	.280	.119	-.163	-.044
	BINOMIAL ERROR	7	.219	.034	.253	.040	.014	.054	.179	.020	.199
TABLE28	BAYES	7	.219	.034	.253	.091	.010	.101	.128	.024	.152
	BAYES	8	.181	.055	.236	.060	.038	.098	.121	.017	.138
	EMPIRICAL BEST	9	.143	.089	.232	.020	.406	.426	.123	-.317	-.194
	PROP CORRECT	7	.249	.004	.253	.122	.093	.215	.127	-.089	.038
	PROP CORRECT	8	.249	.008	.257	.062	.218	.280	.187	-.210	-.023
	BINOMIAL ERROR	4	.257	0	.257	0	0	0	.257	0	.257
	BAYES	7	.249	.004	.253	.047	.003	.050	.202	.001	.203
	BAYES	8	.249	.008	.257	.045	.005	.050	.204	.003	.207
	EMPIRICAL BEST	7	.249	.004	.253	.122	.093	.215	.127	-.089	.038

Table A (cont): 10 Round Easy Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
TABLE35	PROP CORRECT	7	.215	.038	.253	.122	.093	.215	.093	-.055	.038
	PROP CORRECT	8	.156	.068	.224	.062	.218	.280	.094	-.150	-.056
	BINOMIAL ERROR	6	.236	.021	.257	.044	.012	.056	.192	.089	.201
	BAYES	7	.215	.038	.253	.128	.014	.142	.087	.024	.111
	BAYES	8	.156	.068	.224	.082	.057	.139	.074	.011	.085
	EMPIRICAL BEST	8	.156	.068	.224	.062	.218	.280	.094	-.150	-.056
TABLE36	PROP CORRECT	7	.194	.051	.245	.122	.093	.215	.072	-.042	.030
	PROP CORRECT	8	.152	.097	.249	.062	.218	.280	.090	-.121	-.031
	BINOMIAL ERROR	7	.194	.051	.245	.049	.031	.080	.145	.020	.165
	BAYES	7	.194	.051	.245	.122	.018	.140	.072	.033	.105
	BAYES	8	.152	.097	.249	.076	.061	.137	.076	.036	.112
	EMPIRICAL BEST	5	.232	.008	.240	.225	.008	.233	.007	0	.007
TABLE37	PROP CORRECT	7	.228	.004	.232	.122	.093	.215	.106	-.089	.017
	PROP CORRECT	8	.198	.038	.236	.062	.218	.280	.136	-.180	-.044
	BINOMIAL ERROR	6	.241	0	.241	.021	.002	.023	.220	-.002	.218
	BAYES	7	.228	.004	.232	.095	.006	.101	.133	-.002	.131
	BAYES	8	.198	.038	.236	.062	.037	.099	.136	.001	.137
	EMPIRICAL BEST	7	.228	.004	.232	.122	.093	.215	.106	-.089	.017
TABLE38	PROP CORRECT	7	.253	0	.253	.122	.093	.215	.131	-.093	.038
	PROP CORRECT	8	.245	.004	.249	.062	.218	.280	.183	-.214	-.031
	BINOMIAL ERROR	4	.257	0	.257	0	0	0	.257	0	.257
	BAYES	7	.253	0	.253	.050	0	.050	.203	0	.203
	BAYES	8	.245	.004	.249	.043	.006	.049	.202	-.002	.200
	EMPIRICAL BEST	8	.245	.004	.249	.062	.218	.280	.183	-.214	-.031
	EMPIRICAL BEST	9	.232	.017	.249	.020	.406	.426	.212	-.389	-.177

Table A (cont): 10 Round Easy Subtests and 240 Round Criterion



SUBTEST	MODEL	SCORE	OBSERVED MISCLASSIFICATION			EXPECTED MISCLASSIFICATION			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
HARD21	PROP CORRECT	14	.025	.291	.316	.097	.082	.179	-.072	.209	.137
	PROP CORRECT	15	.013	.380	.393	.057	.152	.209	-.044	.228	.184
	BINOMIAL ERROR	15	.013	.380	.393	.088	.092	.180	-.075	.288	.213
	BAYES	14	.025	.291	.316	.116	.056	.172	-.091	.235	.144
	BAYES	15	.013	.380	.393	.058	.111	.169	-.045	.269	.224
	EMPIRICAL BEST	9	.152	.034	.186	.246	0	.246	-.094	.034	-.060
HARD22	PROP CORRECT	14	.017	.342	.359	.097	.082	.179	-.080	.260	.180
	PROP CORRECT	15	.004	.451	.455	.057	.152	.209	-.053	.299	.246
	BINOMIAL ERROR	14	.017	.342	.359	.073	.068	.141	-.056	.274	.218
	BAYES	14	.017	.342	.359	.104	.055	.159	-.087	.287	.200
	BAYES	15	.004	.451	.455	.041	.115	.156	-.037	.336	.299
	EMPIRICAL BEST	8	.143	.025	.168	.253	0	.253	-.110	.025	-.085
HARD23	PROP CORRECT	14	.017	.295	.312	.097	.082	.179	-.080	.223	.133
	PROP CORRECT	15	.004	.384	.388	.057	.152	.209	-.053	.232	.179
	BINOMIAL ERROR	14	.017	.295	.312	.072	.074	.146	-.055	.221	.166
	BAYES	14	.017	.295	.312	.104	.061	.165	-.087	.234	.147
	BAYES	15	.004	.384	.388	.052	.110	.162	-.048	.274	.226
	EMPIRICAL BEST	9	.135	.046	.181	.246	0	.246	-.111	.046	-.065

Table A (cont): 20 Round Hard Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
HARD24	PROP CORRECT	14	.008	.308	.316	.097	.082	.179	-.089	.226	.137
	PROP CORRECT	15	.004	.405	.409	.057	.152	.209	-.053	.253	.200
	BINOMIAL ERROR	14	.008	.308	.316	.070	.062	.132	-.062	.246	.184
	BAYES	14	.008	.308	.316	.103	.052	.155	-.095	.256	.161
	BAYES	15	.004	.405	.409	.051	.101	.152	-.047	.304	.257
	EMPIRICAL BEST	10	.097	.072	.169	.233	.002	.235	-.136	.070	-.066
HARD25	PROP CORRECT	14	.068	.274	.342	.097	.082	.179	-.029	.192	.163
	PROP CORRECT	15	.046	.367	.413	.057	.152	.209	-.011	.215	.204
	BINOMIAL ERROR	15	.046	.367	.413	.065	.106	.171	-.019	.261	.242
	BAYES	14	.068	.274	.342	.114	.052	.166	-.046	.222	.176
	BAYES	15	.046	.367	.413	.056	.107	.163	-.010	.260	.250
	EMPIRICAL BEST	8	.211	.008	.219	.253	0	.253	-.042	.008	-.034
HARD26	EMPIRICAL BEST	9	.177	.042	.219	.246	0	.246	-.069	.042	-.027
	PROP CORRECT	14	.042	.312	.354	.097	.082	.179	-.055	.230	.175
	PROP CORRECT	15	.030	.380	.410	.057	.152	.209	-.027	.228	.201
	BINOMIAL ERROR	15	.030	.380	.410	.073	.099	.172	-.043	.281	.238
	BAYES	14	.042	.312	.354	.098	.067	.165	-.056	.245	.189
	BAYES	15	.030	.380	.410	.057	.106	.163	-.027	.274	.247
	EMPIRICAL BEST	10	.160	.072	.232	.233	.002	.235	-.073	.070	-.003

Table A (cont): 20 Round Hard Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			MISCLASSIFICATION			MISCLASSIFICATION			FP	FN	TOT
EASY22	PROP CORRECT	14	.207	.008	.215	25.9	.097	.082	.110	-.074	.036
	PROP CORRECT	15	.181	.021	.202	8.62	.057	.152	.124	-.131	-.007
	BINOMIAL ERROR	13	.219	0	.219	UND	.019	.008	.200	-.008	.192
	BAYES	14	.207	.008	.215	25.9	.067	.010	.140	-.002	.138
	BAYES	15	.181	.021	.202	8.62	.048	.028	.133	-.007	.126
	EMPIRICAL BEST	16	.152	.034	.186	4.47	.028	.251	.124	-.217	-.093
EASY22	PROP CORRECT	14	.194	.004	.198	48.5	.097	.082	.097	-.078	.019
	PROP CORRECT	15	.173	.013	.186	13.3	.057	.152	.116	-.139	-.023
	BINOMIAL ERROR	12	.245	0	.245	UND	.021	.003	.224	-.003	.221
	BAYES	14	.194	.004	.198	48.5	.064	.016	.130	-.012	.118
	BAYES	15	.173	.013	.186	13.3	.049	.030	.124	-.017	.107
	EMPIRICAL BEST	16	.131	.025	.156	5.24	.028	.251	.103	-.226	-.123
EASY23	PROP CORRECT	14	.207	.004	.211	51.8	.097	.082	.110	-.078	.032
	PROP CORRECT	15	.181	.008	.189	22.6	.057	.152	.124	-.144	-.020
	BINOMIAL ERROR	12	.241	0	.241	UND	.019	.002	.222	-.002	.220
	BAYES	14	.207	.004	.211	51.8	.069	.012	.138	-.008	.130
	BAYES	15	.181	.008	.189	22.6	.054	.026	.127	-.018	.109
	EMPIRICAL BEST	17	.099	.055	.144	1.62	.011	.373	.078	-.318	-.240

Table A (cont): 20 Round Easy Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED MISCLASSIFICATION			EXPECTED MISCLASSIFICATION			DIFFERENCE			
			FP	FN	TOT	FP/FN	FP	FN	TOT	FP	FN	TOT
EASY24	PROP CORRECT	14	.198	.017	.215	11.6	.097	.082	.179	.101	-.065	.036
	PROP CORRECT	15	.169	.017	.186	9.94	.057	.152	.209	.112	-.135	-.023
	BINOMIAL ERROR	13	.232	0	.232	UND	.023	.013	.036	.209	-.013	.196
	BAYES	14	.198	.017	.215	11.6	.064	.015	.079	.134	.002	.136
	BAYES	15	.169	.017	.186	9.94	.049	.029	.078	.120	-.012	.108
	EMPIRICAL BEST	16	.139	.025	.164	5.56	.028	.251	.279	.111	-.226	-.115
EASY25	PROP CORRECT	14	.232	.013	.245	17.8	.097	.082	.179	.135	-.069	.066
	PROP CORRECT	15	.211	.021	.232	8.79	.057	.152	.209	.154	-.131	.023
	BINOMIAL ERROR	12	.257	0	.257	UND	.013	0	.013	.244	0	.244
	BAYES	14	.232	.013	.245	17.8	.057	.010	.067	.175	.003	.178
	BAYES	15	.211	.021	.232	8.79	.042	.024	.066	.169	-.003	.166
	EMPIRICAL BEST	17	.148	.063	.211	2.35	.011	.373	.384	.137	-.310	-.173
EASY26	PROP CORRECT	14	.253	.004	.257	63.3	.097	.082	.179	.156	-.078	.078
	PROP CORRECT	15	.228	.021	.249	10.9	.057	.152	.209	.171	-.131	.040
	BINOMIAL ERROR	11	.257	0	.257	UND	.003	0	.003	.254	0	.254
	BAYES	14	.253	.004	.257	63.3	.067	.002	.069	.186	.002	.188
	BAYES	15	.228	.021	.249	10.9	.045	.023	.068	.183	-.002	.181
	EMPIRICAL BEST	13	.152	.055	.207	2.76	.011	.373	.384	.141	-.318	-.177

Table A (cont): 20 Round Easy Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE	
			FP	FN	TOT	FP	FN	TOT	FP	FN
MIX201	PROP CORRECT	14	.093	.097	.190	.097	.082	.179	-.004	.015
	PROP CORRECT	15	.059	.207	.266	.057	.152	.209	.002	.055
	BINOMIAL ERROR	14	.093	.097	.190	.062	.077	.139	.031	.020
	BAYES	14	.093	.097	.190	.143	.046	.189	-.050	.051
	BAYES	15	.059	.207	.266	.069	.116	.185	-.010	.091
	EMPIRICAL BEST	12	.165	.021	.186	.180	.016	.196	-.015	.005
MIX202	PROP CORRECT	14	.105	.084	.189	.097	.082	.179	.008	.002
	PROP CORRECT	15	.072	.160	.232	.057	.152	.209	.015	.008
	BINOMIAL ERROR	14	.105	.084	.189	.058	.076	.134	.047	.008
	BAYES	14	.105	.084	.189	.147	.040	.187	-.042	.044
	BAYES	15	.072	.160	.232	.091	.093	.184	-.019	.067
	EMPIRICAL BEST	12	.169	.013	.182	.180	.016	.196	-.011	.003
MIX203	PROP CORRECT	14	.080	.110	.190	.097	.082	.179	-.017	.028
	PROP CORRECT	15	.034	.198	.232	.057	.152	.209	-.023	.046
	BINOMIAL ERROR	14	.080	.110	.190	.064	.084	.148	.016	.086
	BAYES	14	.080	.110	.190	.136	.045	.181	-.056	.065
	BAYES	15	.034	.198	.232	.067	.111	.178	-.033	.087
	EMPIRICAL BEST	12	.152	.025	.177	.180	.016	.196	-.028	.009
MIX204	PROP CORRECT	14	.080	.118	.198	.097	.082	.179	-.017	.036
	PROP CORRECT	15	.046	.198	.244	.057	.152	.209	-.011	.046
	BINOMIAL ERROR	14	.080	.118	.198	.069	.091	.160	.011	.027
	BAYES	14	.080	.118	.198	.144	.048	.192	-.064	.070
	BAYES	15	.046	.198	.244	.086	.103	.189	-.040	.095
	EMPIRICAL BEST	12	.143	.021	.164	.180	.016	.196	-.037	.005

Table A (cont.): 20 Round Mix Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
MIX205	PROP CORRECT	14	.089	.068	.157	.097	.082	.179	-.008	-.014	-.022
	PROP CORRECT	15	.059	.143	.202	.057	.152	.209	.002	-.009	-.007
	BINOMIAL ERROR	13	.127	.051	.178	.074	.061	.135	.053	-.010	.043
	BAYES	14	.089	.068	.157	.145	.037	.182	-.056	.031	-.025
	BAYES	15	.059	.143	.202	.091	.088	.179	-.032	.055	.023
	EMPIRICAL BEST	14	.089	.068	.157	.097	.082	.179	-.008	-.014	-.022
MIX206	PROP CORRECT	14	.080	.089	.169	.097	.082	.179	-.017	.007	-.010
	PROP CORRECT	15	.059	.143	.202	.057	.152	.209	.002	-.009	-.007
	BINOMIAL ERROR	14	.080	.089	.169	.047	.070	.117	.033	.019	.052
	BAYES	14	.080	.089	.169	.114	.041	.155	-.034	.048	.014
	BAYES	15	.059	.143	.202	.075	.078	.153	-.016	.065	.049
	EMPIRICAL BEST	13	.122	.046	.168	.141	.039	.180	-.019	.007	-.012
MIX207	PROP CORRECT	14	.080	.105	.185	.097	.082	.179	-.017	.023	.006
	PROP CORRECT	15	.046	.173	.219	.057	.152	.209	-.011	.021	.010
	BINOMIAL ERROR	14	.080	.105	.185	.060	.084	.144	.020	.021	.041
	BAYES	14	.080	.105	.185	.133	.049	.182	-.053	.056	.003
	BAYES	15	.046	.173	.219	.081	.098	.179	-.035	.075	.040
	EMPIRICAL BEST	13	.110	.034	.144	.141	.039	.180	-.031	-.005	-.036
MIX208	PROP CORRECT	14	.055	.118	.173	.097	.082	.179	-.042	.036	-.006
	PROP CORRECT	15	.025	.186	.211	.057	.152	.209	-.032	.034	.002
	BINOMIAL ERROR	14	.055	.118	.173	.050	.089	.139	.005	.029	.034
	BAYES	14	.055	.118	.173	.108	.049	.157	-.053	.069	.016
	BAYES	15	.025	.186	.211	.058	.096	.154	-.033	.090	.057
	EMPIRICAL BEST	13	.089	.068	.157	.141	.039	.180	-.052	.029	-.023

Table A (cont): 20 Round Mix Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
MIX209	PROP CORRECT	14	.131	.076	.207	.097	.082	.179	.034	-.006	.028
	PROP CORRECT	15	.101	.135	.236	.057	.152	.209	.044	-.017	.027
	BINOMIAL ERROR	14	.131	.076	.207	.042	.060	.102	.089	.016	.105
	BAYES	14	.131	.076	.207	.121	.031	.152	.010	.045	.055
	BAYES	15	.101	.135	.236	.076	.074	.150	.025	.061	.086
MIX210	EMPIRICAL BEST	14	.131	.076	.207	.097	.082	.179	.034	-.006	.028
	PROP CORRECT	14	.139	.093	.232	.097	.082	.179	.042	.011	.053
	PROP CORRECT	15	.118	.135	.253	.057	.152	.209	.061	-.017	.044
	BINOMIAL ERROR	13	.177	.046	.223	.058	.046	.104	.119	0	.119
	BAYES	14	.139	.093	.232	.126	.043	.169	.013	.050	.063
MIX211	BAYES	15	.118	.135	.253	.094	.074	.168	.024	.061	.085
	EMPIRICAL BEST	13	.177	.046	.223	.141	.039	.180	.036	.007	.043
	PROP CORRECT	14	.160	.110	.270	.097	.082	.179	.063	.028	.091
	PROP CORRECT	15	.118	.165	.283	.057	.152	.209	.061	.013	.074
	BINOMIAL ERROR	13	.198	.042	.240	.117	.036	.153	.081	.006	.087
MIX212	BAYES	14	.160	.110	.270	.139	.044	.183	.021	.066	.087
	BAYES	15	.118	.165	.283	.089	.091	.180	.029	.074	.103
	EMPIRICAL BEST	13	.198	.042	.240	.141	.039	.180	.057	.003	.060
	PROP CORRECT	14	.127	.089	.216	.097	.082	.179	.030	.007	.037
	PROP CORRECT	15	.101	.156	.257	.057	.152	.209	.044	.004	.048
MIX221	BINOMIAL ERROR	14	.127	.089	.216	.075	.060	.135	.052	.029	.081
	BAYES	14	.127	.089	.216	.123	.036	.159	.004	.053	.057
	BAYES	15	.101	.156	.257	.075	.081	.156	.026	.075	.101
	EMPIRICAL BEST	13	.152	.046	.198	.141	.039	.180	.011	.007	.018
	PROP CORRECT	14	.127	.089	.216	.097	.082	.179	.030	.007	.037

Table A (cont): 20 Round Mix Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
HARD41	PROP CORRECT	27	.008	.304	.312	.100	.040	.140	-.092	.264	.172
	PROP CORRECT	28	0	.354	.354	.073	.066	.139	-.073	.288	.215
	PROP CORRECT	29	0	.401	.401	.049	.103	.152	-.049	.298	.249
	BINOMIAL ERROR	29	0	.401	.401	.062	.060	.122	-.062	.341	.279
	BAYES	28	0	.354	.354	.069	.050	.119	-.069	.304	.235
	BAYES	29	0	.401	.401	.045	.073	.118	-.045	.328	.283
HARD42	EMPIRICAL BEST	18	.135	.030	.165	.252	0	.252	-.117	.030	-.087
	PROP CORRECT	27	0	.278	.278	.100	.040	.140	-.100	.238	.138
	PROP CORRECT	28	0	.312	.312	.073	.066	.139	-.073	.246	.173
	PROP CORRECT	29	0	.367	.367	.049	.103	.152	-.049	.264	.215
	BINOMIAL ERROR	29	0	.367	.367	.067	.059	.126	-.067	.308	.241
	BAYES	28	0	.312	.312	.082	.042	.124	-.082	.270	.188
HARD43	BAYES	29	0	.367	.367	.054	.069	.123	-.054	.298	.244
	EMPIRICAL BEST	19	.110	.034	.144	.248	0	.248	-.138	.034	-.104
	PROP CORRECT	27	.059	.245	.304	.100	.040	.140	-.041	.205	.164
	PROP CORRECT	28	.038	.295	.333	.073	.066	.139	-.035	.229	.194
	PROP CORRECT	29	.021	.333	.354	.049	.103	.152	-.028	.230	.202
	BINOMIAL ERROR	29	.021	.333	.354	.061	.074	.135	-.040	.259	.219
HARD43	BAYES	28	.038	.295	.333	.082	.051	.133	-.044	.244	.200
	BAYES	29	.021	.333	.354	.054	.078	.132	-.033	.255	.222
	EMPIRICAL BEST	24	.076	.139	.215	.183	.006	.189	-.107	.133	.026

Table A (cont): 40 Round Hard Subtests and 240 Round Criterion



SUBTEST	MODEL	SCORE	OBSERVED MISCLASSIFICATION				EXPECTED MISCLASSIFICATION				DIFFERENCE	
			FP	FN	TOT	FP/FN	FP	FN	TOT	FP	FN	TOT
EASY41	PROP CORRECT	27	.228	0	.228	UND	.100	.040	.140	.128	-.040	.088
	PROP CORRECT	28	.203	.004	.207	50.8	.073	.066	.139	.130	-.062	.068
	PROP CORRECT	29	.181	.004	.185	45.3	.049	.103	.152	.132	-.099	.033
	BINOMIAL ERROR	26	.232	0	.232	UND	.014	.003	.017	.218	-.003	.215
	BAYES	28	.203	.004	.207	50.8	.031	.013	.044	.172	-.009	.163
	BAYES	29	.181	.004	.185	45.3	.020	.023	.043	.161	-.019	.142
EASY42	EMPIRICAL BEST	33	.110	.025	.135	4.40	.004	.352	.356	.106	-.327	-.221
	PROP CORRECT	27	.215	0	.215	UND	.100	.040	.140	.115	-.040	.075
	PROP CORRECT	28	.211	0	.211	UND	.073	.066	.139	.138	-.066	.072
	PROP CORRECT	29	.198	0	.198	UND	.049	.103	.152	.149	-.103	.046
	BINOMIAL ERROR	27	.215	0	.215	UND	.010	.009	.019	.205	-.009	.196
	BAYES	28	.211	0	.211	UND	.043	.007	.050	.168	-.007	.161
EASY43	BAYES	29	.198	0	.198	UND	.037	.013	.050	.161	-.013	.148
	EMPIRICAL BEST	33	.089	.046	.135	1.93	.004	.352	.356	.085	-.306	-.221
	PROP CORRECT	27	.253	.004	.257	63.3	.100	.040	.140	.153	-.036	.117
	PROP CORRECT	28	.241	.004	.245	60.3	.073	.066	.139	.168	-.062	.106
	PROP CORRECT	29	.232	.008	.240	29.0	.049	.103	.152	.183	-.095	.088
	BINOMIAL ERROR	25	.257	0	.257	UND	.007	0	.007	.250	0	.250
EASY43	BAYES	28	.241	.004	.245	60.3	.030	.006	.036	.211	-.002	.209
	BAYES	29	.232	.008	.240	29.0	.024	.012	.036	.208	-.004	.204
	EMPIRICAL BEST	35	.165	.046	.211	3.59	0	.499	.499	.165	-.453	-.288

Table A (cont): 40 Round Easy Subtests and 240 Round Criterion

			OBSERVED				EXPECTED				DIFFERENCE		
			MISCLASSIFICATION				MISCLASSIFICATION				DIFFERENCE		
SUBTEST	MODEL	SCORE	FP	FN	TOT	FP/FN	FP	FN	TOT	FP	FN	TOT	
MIX41	PROP CORRECT	27	.089	.042	.131	2.12	.100	.040	.140	-.011	.002	-.009	
	PROP CORRECT	28	.072	.080	.152	.900	.073	.066	.139	-.001	.014	.013	
	PROP CORRECT	29	.063	.135	.198	.467	.049	.103	.152	.014	.032	.046	
	BINOMIAL ERROR	28	.072	.080	.152	.900	.073	.063	.136	-.001	.017	.016	
	BAYES	28	.072	.080	.152	.900	.102	.043	.145	-.030	.037	.007	
	BAYES	29	.063	.135	.198	.467	.070	.074	.144	-.007	.061	.054	
	EMPIRICAL BEST	27	.089	.042	.131	2.12	.100	.040	.144	-.011	.002	.009	
MIX42	PROP CORRECT	27	.063	.051	.114	1.24	.100	.040	.140	-.037	.011	-.026	
	PROP CORRECT	28	.051	.084	.135	.607	.073	.066	.139	-.022	.018	-.004	
	PROP CORRECT	29	.042	.152	.194	.276	.049	.103	.152	-.007	.049	.042	
	BINOMIAL ERROR	28	.051	.084	.135	.607	.085	.054	.139	-.034	.030	-.004	
	BAYES	28	.051	.084	.135	.607	.106	.040	.146	-.055	.044	-.011	
	BAYES	29	.042	.152	.194	.276	.067	.077	.144	-.025	.075	.050	
	EMPIRICAL BEST	27	.063	.051	.114	1.24	.100	.040	.140	-.037	.011	-.026	
MIX43	PROP CORRECT	27	.097	.046	.143	2.11	.100	.040	.140	-.003	.006	.003	
	PROP CORRECT	28	.072	.068	.140	1.06	.073	.066	.139	-.001	.002	.001	
	PROP CORRECT	29	.059	.097	.156	.608	.049	.103	.152	.010	.006	.004	
	BINOMIAL ERROR	28	.072	.068	.140	1.06	.062	.044	.106	.010	.024	.034	
	BAYES	28	.072	.068	.140	1.06	.086	.033	.119	-.014	.035	.021	
	BAYES	29	.059	.097	.156	.608	.065	.054	.119	-.006	.043	.037	
	EMPIRICAL BEST	26	.105	.034	.139	3.09	.129	.022	.151	-.024	.012	-.012	
MIX44	PROP CORRECT	27	.059	.046	.105	1.28	.100	.040	.140	-.041	.006	-.035	
	PROP CORRECT	28	.051	.084	.135	.607	.073	.066	.139	-.022	.018	-.004	
	PROP CORRECT	29	.038	.143	.181	.266	.049	.103	.152	-.011	.040	.029	
	BINOMIAL ERROR	28	.051	.084	.135	.607	.076	.048	.124	-.025	.086	.011	
	BAYES	28	.051	.084	.135	.607	.096	.039	.135	-.045	.045	0	
	BAYES	29	.038	.143	.181	.266	.059	.074	.133	-.021	.069	.048	
	EMPIRICAL BEST	27	.059	.046	.105	1.28	.100	.040	.140	-.041	.006	-.035	

File A (cont): 40 Round Mix Subtests and 240 Round Criterion

**Table A (cont): 40 Round Mix Subtests and 240 Round Criterion**

SUBTEST	MODEL	SCORE	OBSERVED				EXPECTED				DIFFERENCE	
			FP	FN	TOT	FP/FN	FP	FN	TOT	FP	FN	TOT
MIX45	PROP CORRECT	27	.152	.042	.194	3.62	.100	.040	.140	.052	.002	.054
	PROP CORRECT	28	.127	.068	.195	1.87	.073	.066	.139	.054	.002	.056
	PROP CORRECT	29	.114	.105	.219	1.09	.049	.103	.152	.065	.002	.067
	BINOMIAL ERROR	27	.152	.042	.194	3.62	.074	.023	.097	.078	.019	.097
	BAYES	28	.127	.068	.195	1.87	.086	.033	.119	.041	.035	.076
	BAYES	29	.114	.105	.219	1.09	.060	.058	.118	.054	.047	.101
	EMPIRICAL BEST	27	.152	.042	.194	3.62	.100	.040	.140	.052	.002	.054
MIX46	PROP CORRECT	27	.152	.055	.207	2.76	.100	.040	.140	.052	.015	.067
	PROP CORRECT	28	.127	.093	.220	1.37	.073	.066	.139	.054	.027	.081
	PROP CORRECT	29	.105	.114	.219	.921	.049	.103	.152	.056	.011	.067
	BINOMIAL ERROR	27	.152	.055	.207	2.76	.079	.037	.116	.073	.018	.091
	BAYES	28	.127	.093	.220	1.37	.089	.044	.133	.038	.049	.087
	BAYES	29	.105	.114	.219	.921	.068	.065	.133	.037	.049	.086
	EMPIRICAL BEST	27	.152	.055	.207	2.76	.100	.040	.140	.052	.015	.067

Table A (cont): 40 Round Mix Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
REP1	PROP CORRECT	54	.068	.034	.102	.084	.024	.108	-.016	.010	-.006
	PROP CORRECT	55	.063	.051	.114	.067	.035	.102	-.004	.016	.012
	PROP CORRECT	56	.051	.063	.114	.052	.050	.102	-.001	.013	.012
	BINOMIAL ERROR	56	.051	.063	.114	.065	.039	.104	-.014	.024	.010
	BAYES	56	.051	.063	.114	.085	.031	.116	-.034	.032	-.002
	BAYES	57	.034	.089	.123	.064	.052	.116	-.030	.037	.007
REP2	EMPIRICAL BEST	53	.068	.021	.089	.103	.016	.119	-.035	.005	-.030
	PROP CORRECT	54	.055	.042	.097	.084	.024	.108	-.029	.018	-.011
	PROP CORRECT	55	.038	.055	.093	.067	.035	.102	-.029	.020	-.009
	PROP CORRECT	56	.030	.055	.085	.052	.050	.102	-.022	.005	-.017
	BINOMIAL ERROR	56	.030	.055	.085	.041	.034	.085	-.011	.021	.010
	BAYES	56	.030	.055	.085	.054	.029	.083	-.024	.026	.002
REP3	BAYES	57	.030	.080	.110	.041	.042	.083	-.011	.038	.027
	EMPIRICAL BEST	56	.030	.055	.085	.052	.050	.102	-.022	.005	-.017
	PROP CORRECT	54	.143	.051	.194	.084	.024	.108	.059	.027	.086
	PROP CORRECT	55	.139	.063	.202	.067	.035	.102	.072	.028	.100
	PROP CORRECT	56	.122	.084	.206	.053	.050	.102	.070	.034	.104
	BINOMIAL ERROR	55	.139	.063	.202	.050	.031	.081	.089	.032	.121
REP3	BAYES	56	.122	.084	.206	.054	.038	.092	.068	.046	.114
	BAYES	57	.105	.093	.198	.041	.051	.092	.064	.042	.106
	EMPIRICAL BEST	54	.143	.051	.194	.084	.024	.108	.059	.027	.086
	EMPIRICAL BEST	54	.143	.051	.194	.084	.024	.108	.059	.027	.086

Table A (cont): 80 Round Subtests and 240 Round Criterion

SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY	SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY
TABLE11	PROP CORRECT	11.634	-22.783	TABLE15	PROP CORRECT	5.480	16.217
	BINOMIAL ERROR	6.370	-22.795		BINOMIAL ERROR	3.062	16.213
	BAYES	7.199	-24.334		BAYES	3.476	2.949
TABLE12	PROP CORRECT	13.690	-34.583	TABLE16	PROP CORRECT	6.190	13.017
	BINOMIAL ERROR	8.374	-34.594		BINOMIAL ERROR	3.233	13.007
	BAYES	8.682	-32.588		BAYES	2.955	0.710
TABLE13	PROP CORRECT	20.120	-49.783	TABLE17	PROP CORRECT	7.710	32.117
	BINOMIAL ERROR	14.421	-49.792		BINOMIAL ERROR	6.298	32.104
	BAYES	12.436	-43.221		BAYES	2.977	14.071
TABLE14	PROP CORRECT	11.474	-27.383	TABLE18	PROP CORRECT	12.533	47.517
	BINOMIAL ERROR	6.498	-27.395		BINOMIAL ERROR	11.965	47.502
	BAYES	7.072	-27.552		BAYES	5.190	24.844
TABLE21	PROP CORRECT	8.625	-20.483	TABLE25	PROP CORRECT	6.157	18.017
	BINOMIAL ERROR	4.263	-20.500		BINOMIAL ERROR	3.647	18.002
	BAYES	5.225	-22.725		BAYES	2.665	4.208
TABLE22	PROP CORRECT	11.639	-28.483	TABLE26	PROP CORRECT	7.327	12.717
	BINOMIAL ERROR	6.721	-28.500		BINOMIAL ERROR	4.232	12.705
	BAYES	7.094	-28.321		BAYES	3.279	0.500
TABLE23	PROP CORRECT	22.832	-54.983	TABLE27	PROP CORRECT	9.481	32.417
	BINOMIAL ERROR	17.100	-54.996		BINOMIAL ERROR	7.663	32.401
	BAYES	13.898	-46.859		BAYES	3.863	14.281
TABLE24	PROP CORRECT	11.938	-23.583	TABLE28	PROP CORRECT	12.744	48.517
	BINOMIAL ERROR	6.907	-23.598		BINOMIAL ERROR	12.323	48.494
	BAYES	7.016	-24.893		BAYES	5.267	25.544

Table B: Average Per Test Sum of Squared and Absolute Discrepancies Between Estimated True Scores and Criterion True Scores: 10 Round Subtests and 240 Round Criterion

SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY	SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY
TABLE31	PROP CORRECT	12.501	-27.283	TABLE35	PROP CORRECT		
	BINOMIAL ERROR	7.343	-27.297		BINOMIAL ERROR		
	BAYES	7.534	-27.482		BAYES		
TABLE32	PROP CORRECT	10.534	-18.883	TABLE36	PROP CORRECT		
	BINOMIAL ERROR	5.496	-18.894		BINOMIAL ERROR		
	BAYES	6.458	-21.606		BAYES		
TABLE33	PROP CORRECT	15.872	-35.383	TABLE37	PROP CORRECT		
	BINOMIAL ERROR	10.156	-35.398		BINOMIAL ERROR		
	BAYES	9.783	-33.148		BAYES		
TABLE34	PROP CORRECT	10.220	-12.883	TABLE38	PROP CORRECT		
	BINOMIAL ERROR	5.311	-12.893		BINOMIAL ERROR		
	BAYES	6.029	-17.408		BAYES		
							189

Table B (cont): 10 Round Subtests and 240 Round Criterion

SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY	SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY
HARD21	PROP CORRECT	8.161	-30.183	HARD23	PROP CORRECT	8.190	-29.933
	BINOMIAL ERROR	5.901	-30.132		BINOMIAL ERROR	4.490	-22.362
	BAYES	6.580	-29.787		BAYES	6.409	-29.582
HARD25	PROP CORRECT	6.806	-21.833	HARD26	PROP CORRECT	7.350	-25.383
	BINOMIAL ERROR	4.537	-11.846		BINOMIAL ERROR	5.062	-25.393
	BAYES	5.340	-22.914		BAYES	5.884	-25.836
EASY21	PROP CORRECT	5.230	26.917	EASY22	PROP CORRECT	5.121	27.517
	BINOMIAL ERROR	4.605	26.902		BINOMIAL ERROR	4.683	27.503
	BAYES	3.005	17.218		BAYES	2.947	17.712
EASY23	PROP CORRECT	5.237	28.867	EASY24	PROP CORRECT	5.208	26.967
	BINOMIAL ERROR	4.842	28.858		BINOMIAL ERROR	4.570	26.954
	BAYES	2.928	18.823		BAYES	2.956	17.259
EASY25	PROP CORRECT	6.912	33.567	EASY26	PROP CORRECT	6.962	34.417
	BINOMIAL ERROR	6.499	33.557		BINOMIAL ERROR	6.726	34.401
	BAYES	4.057	22.693		BAYES	4.087	23.392

Table B (cont): 20 Round Hard and Easy Subtests and 240 Round Criterion

SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY	SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY
MIX201	PROP CORRECT BINOMIAL ERROR BAYES	2.856 1.489 2.189	-1.983 -1.994 -6.573	MIX202	PROP CORRECT BINOMIAL ERROR BAYES	2.334 1.281 1.839	-1.283 -1.296 -5.997
MIX203	PROP CORRECT BINOMIAL ERROR BAYES	2.843 1.464 2.216	-4.133 -4.144 -8.343	MIX204	PROP CORRECT BINOMIAL ERROR BAYES	2.817 1.510 2.291	-5.433 -5.446 -9.413
MIX205	PROP CORRECT BINOMIAL ERROR BAYES	2.086 1.150 1.630	-0.933 -0.943 -5.708	MIX206	PROP CORRECT BINOMIAL ERROR BAYES	2.721 1.334 1.865	-0.133 -0.145 -5.050
MIX207	PROP CORRECT BINOMIAL ERROR BAYES	2.598 1.322 2.015	-3.533 -3.554 -7.849	MIX208	PROP CORRECT BINOMIAL ERROR BAYES	3.141 1.587 2.260	-3.333 -3.347 -7.684
MIX209	PROP CORRECT BINOMIAL ERROR BAYES	3.516 2.046 2.383	5.917 5.906 -0.069	MIX210	PROP CORRECT BINOMIAL ERROR BAYES	2.812 1.687 2.006	5.817 5.805 -0.152
MIX211	PROP CORRECT BINOMIAL ERROR BAYES	3.022 1.796 2.256	4.467 4.411 -1.263	MIX212	PROP CORRECT BINOMIAL ERROR BAYES	3.120 1.756 2.188	4.567 4.556 -1.181

Table B (cont): 20 Round Mix Subtests and 240 Round Criterion



SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY	SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY
HARD41	PROP CORRECT			HARD42	PROP CORRECT		
	BINOMIAL ERROR				BINOMIAL ERROR		
	BAYES				BAYES		
HARD43	PROP CORRECT	7.817	-33.633	EASY42	PROP CORRECT	4.655	27.917
	BINOMIAL ERROR	6.706	-33.644		BINOMIAL ERROR	4.457	27.910
	BAYES	6.995	-33.082		BAYES	3.382	22.500
EASY41	PROP CORRECT	5.672	-23.608	MIX42	PROP CORRECT	1.627	-4.783
	BINOMIAL ERROR	4.570	-23.620		BINOMIAL ERROR	1.101	-4.791
	BAYES	5.061	-24.029		BAYES	1.475	-7.029
EASY43	PROP CORRECT	4.346	27.217	MIX44	PROP CORRECT	1.637	-3.433
	BINOMIAL ERROR	4.253	27.208		BINOMIAL ERROR	1.036	-3.444
	BAYES	3.169	21.868		BAYES	1.397	-5.810
EASY45	PROP CORRECT	6.470	33.992	MIX46	PROP CORRECT	1.718	4.517
	BINOMIAL ERROR	6.386	33.978		BINOMIAL ERROR	1.285	4.504
	BAYES	4.843	27.986		BAYES	1.428	1.369
MIX41	PROP CORRECT	1.383	-1.633				
	BINOMIAL ERROR	0.953	-1.646				
	BAYES	1.229	-4.185				
MIX43	PROP CORRECT	1.367	-0.533				
	BINOMIAL ERROR	0.860	-0.547				
	BAYES	1.132	-3.191				
MIX45	PROP CORRECT	1.960	5.867				
	BINOMIAL ERROR	1.436	5.855				
	BAYES	1.574	2.588				

Table B (cont): 40 Round Hard, Easy, and Mix Subtests and 240 Round Criterion

SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY	SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY
REP1	PROP CORRECT	0.871	-3.208	REP2	PROP CORRECT	0.831	-1.983
	BINOMIAL ERROR	0.699	-3.221		BINOMIAL ERROR	0.610	-1.995
	BAYES	0.837	-4.469		BAYES	0.755	-3.306
REP3	PROP CORRECT	1.358	5.192				
	BINOMIAL ERROR	1.152	5.181				
	BAYES	1.209	3.503				

Table B (cont): 80 Round Subtests and 240 Round Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP/FN	FP	FN	TOT	FP	FN
TABLE11	PROP CORRECT	7	.249	.068	.317	3.66	.195	.059	.252	.054	.011
	PROP CORRECT	8	.169	.089	.258	1.90	.088	.125	.213	.081	-.036
	BINOMIAL ERROR	8	.169	.089	.258	1.90	.080	.102	.182	.089	-.013
	BAYES	8	.169	.089	.258	1.90	.144	.065	.209	.025	.024
	EMPIRICAL BEST	9	.072	.156	.228	.462	.026	.220	.246	.046	-.064
TABLE12	PROP CORRECT	7	.177	.093	.270	1.90	.195	.059	.252	-.018	.036
	PROP CORRECT	8	.127	.156	.283	.814	.088	.125	.213	.039	.031
	BINOMIAL ERROR	8	.127	.156	.283	.814	.084	.103	.187	.043	.053
	BAYES	8	.127	.156	.283	.814	.116	.077	.193	.011	.079
	EMPIRICAL BEST	9	.042	.219	.261	.192	.026	.220	.246	.016	-.001
TABLE13	PROP CORRECT	7	.165	.127	.292	1.30	.195	.059	.252	-.030	.070
	PROP CORRECT	8	.068	.194	.262	.351	.088	.125	.213	-.020	.069
	BINOMIAL ERROR	8	.068	.194	.262	.351	.073	.090	.163	-.005	.107
	BAYES	8	.068	.194	.262	.351	.086	.083	.169	-.018	.111
	EMPIRICAL BEST	8	.068	.194	.262	.351	.088	.125	.213	-.020	.069
TABLE14	PROP CORRECT	7	.186	.063	.249	2.95	.195	.057	.252	-.009	.006
	PROP CORRECT	8	.127	.110	.237	1.15	.088	.125	.213	.039	-.015
	BINOMIAL ERROR	8	.127	.110	.237	1.15	.073	.113	.186	.054	-.003
	BAYES	8	.127	.110	.237	1.15	.120	.075	.195	.007	.035
	EMPIRICAL BEST	9	.046	.165	.211	.279	.026	.220	.246	.020	-.055

Table C: Recommended Criterion Scores and Observed, Expected, and Observed versus Expected Misclassification Rates: 10 Round Hard Subtests and 120 Round Hard Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
TABLE21	PROP CORRECT	7	.249	.042	.291	.195	.057	.252	.054	-.015	.039
	PROP CORRECT	8	.165	.097	.262	.088	.125	.213	.077	-.028	.049
	BINOMIAL ERROR	8	.165	.097	.262	.076	.129	.205	.089	-.032	.057
	BAYES	8	.165	.097	.262	.138	.079	.217	.027	.018	.045
	EMPIRICAL BEST	9	.076	.169	.245	.026	.220	.246	.050	-.051	-.001
TABLE22	PROP CORRECT	7	.219	.051	.270	.195	.057	.252	.024	-.006	.018
	PROP CORRECT	8	.114	.118	.232	.088	.125	.213	.026	-.007	.019
	BINOMIAL ERROR	8	.114	.118	.232	.067	.125	.192	.047	-.007	.040
	BAYES	8	.114	.118	.232	.066	.111	.085	.003	.033	.036
	EMPIRICAL BEST	8	.114	.118	.232	.066	.125	.213	.026	-.007	.019
TABLE23	PROP CORRECT	7	.089	.122	.211	.195	.057	.252	-.106	.065	-.041
	PROP CORRECT	8	.025	.190	.215	.088	.125	.213	-.063	.065	.002
	BINOMIAL ERROR	8	.025	.190	.215	.050	.080	.130	-.025	.110	.085
	BAYES	8	.025	.190	.215	.060	.079	.139	-.035	.111	.076
	EMPIRICAL BEST	7	.089	.122	.211	.195	.057	.252	-.106	.065	-.041
TABLE24	PROP CORRECT	7	.257	.051	.308	.195	.057	.252	.062	-.006	.056
	PROP CORRECT	8	.181	.089	.270	.088	.125	.213	.093	-.036	.057
	BINOMIAL ERROR	8	.181	.089	.270	.077	.098	.175	.104	-.009	.095
	BAYES	8	.181	.089	.270	.139	.064	.203	.042	.025	.067
	EMPIRICAL BEST	9	.076	.165	.241	.026	.220	.246	.050	-.055	-.005

Table C (cont): 10 Round Hard Subtests and 120 Round Hard Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
TABLE31	PROP CORRECT	7	.245	.042	.287	.195	.057	.252	.050	-.015	.035
	PROP CORRECT	8	.156	.105	.261	.088	.125	.213	.068	-.020	.048
	EMPIRICAL BEST	8	.156	.105	.261	.081	.110	.191	.075	-.005	.070
	BAYES	8	.156	.105	.261	.141	.074	.215	.015	.031	.046
	EMPIRICAL BEST	9	.051	.186	.237	.026	.220	.246	.025	-.034	-.009
TABLE32	PROP CORRECT	7	.287	.063	.350	.195	.057	.252	.092	.006	.098
	PROP CORRECT	8	.203	.097	.300	.088	.125	.213	.115	-.028	.087
	BINOMIAL ERROR	8	.203	.097	.300	.077	.113	.190	.126	-.016	.110
	BAYES	8	.203	.097	.300	.142	.068	.210	.061	.029	.090
	EMPIRICAL BEST	9	.093	.156	.249	.026	.220	.246	.067	-.064	.003
TABLE33	PROP CORRECT	7	.211	.114	.325	.195	.057	.252	.016	.057	.073
	PROP CORRECT	8	.122	.139	.261	.088	.125	.213	.034	.014	.048
	BINOMIAL ERROR	8	.122	.139	.261	.073	.094	.167	.049	.045	.094
	BAYES	8	.122	.139	.261	.113	.070	.183	.009	.069	.078
	EMPIRICAL BEST	9	.046	.203	.249	.026	.220	.246	.020	-.007	.003
TABLE34	PROP CORRECT	7	.359	.068	.427	.195	.057	.252	.164	.011	.175
	PROP CORRECT	8	.266	.101	.367	.088	.125	.213	.178	-.024	.154
	BINOMIAL ERROR	7	.359	.068	.427	.134	.045	.179	.225	.023	.248
	BAYES	8	.266	.101	.367	.170	.066	.236	.096	.035	.131
	EMPIRICAL BEST	9	.101	.148	.249	.026	.220	.246	.075	-.072	.003

Table C (cont): 10 Round Hard Subtests and 120 Round Hard Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
HARD21	PROP CORRECT	14	.173	.063	.236	.130	.053	.183	.043	.010	.053
	PROP CORRECT	15	.101	.093	.194	.072	.094	.166	.029	-.001	.028
	BINOMIAL ERROR	15	.101	.093	.194	.088	.092	.180	.013	.001	.014
	BAYES	15	.101	.093	.194	.107	.081	.188	-.006	.012	.006
	EMPIRICAL BEST	15	.101	.093	.194	.072	.094	.166	.029	-.001	.028
HARD22	PROP CORRECT	14	.105	.055	.160	.130	.053	.183	-.025	.002	-.023
	PROP CORRECT	15	.051	.122	.173	.072	.094	.166	-.021	.028	.007
	BINOMIAL ERROR	14	.105	.055	.160	.073	.068	.141	.032	-.013	.019
	BAYES	15	.051	.122	.173	.072	.090	.162	-.021	.032	.011
	EMPIRICAL BEST	14	.105	.055	.160	.130	.053	.183	-.025	.002	-.023
HARD23	PROP CORRECT	14	.148	.051	.199	.130	.053	.183	.018	-.002	.016
	PROP CORRECT	15	.089	.093	.182	.072	.094	.166	.017	-.001	.016
	BINOMIAL ERROR	14	.146	.051	.199	.072	.074	.146	.076	-.023	.053
	BAYES	15	.089	.093	.182	.090	.085	.175	-.001	.008	.007
	EMPIRICAL BEST	16	.051	.127	.178	.033	.147	.180	.018	-.020	-.002

Table C (cont): 20 Round Hard Subtests and 120 Round Hard Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
HARD24	PROP CORRECT	14	.143	.068	.211	.130	.053	.183	.013	.015	.028
	PROP CORRECT	15	.097	.122	.219	.072	.094	.166	.025	.028	.053
	BINOMIAL ERROR	14	.143	.068	.211	.070	.062	.132	.073	.006	.079
	BAYES	15	.097	.122	.219	.087	.078	.165	.010	.044	.054
	EMPIRICAL BEST	13	.177	.030	.207	.204	.026	.230	-.027	.004	-.023
HARD25	PROP CORRECT	14	.207	.038	.245	.130	.053	.183	.077	-.015	.062
	PROP CORRECT	15	.135	.080	.215	.072	.094	.166	.063	-.014	.049
	BINOMIAL ERROR	15	.135	.080	.215	.065	.106	.171	.070	-.026	.044
	BAYES	15	.135	.080	.215	.096	.085	.181	.039	-.005	.034
	EMPIRICAL BEST	15	.135	.080	.215	.072	.094	.166	.063	-.014	.049
HARD26	PROP CORRECT	14	.194	.089	.283	.130	.053	.183	.064	.036	.100
	PROP CORRECT	15	.143	.118	.261	.072	.094	.166	.071	.024	.095
	BINOMIAL ERROR	15	.143	.118	.261	.073	.099	.172	.070	.019	.089
	BAYES	15	.143	.118	.261	.097	.079	.176	.046	.049	.085
	EMPIRICAL BEST	15	.143	.118	.261	.072	.094	.166	.071	.024	.095

Table C (cont): 20 Round Hard Subtests and 120 Round Hard Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			MISCLASSIFICATION			MISCLASSIFICATION					
			FP	FN	TOT	FP/FN	FP	FN	TOT	FP	FN
HARD41	PROP CORRECT	27	.118	.038	.156	3.11	.120	.028	.148	-.002	.010
	PROP CORRECT	28	.076	.055	.131	1.38	.084	.045	.129	-.008	.010
	PROP CORRECT	29	.055	.080	.135	.688	.055	.069	.124	0	.011
	BINOMIAL ERROR	29	.055	.080	.135	.688	.062	.060	.122	-.007	.020
	BAYES	29	.055	.080	.135	.688	.067	.057	.124	-.012	.023
	EMPIRICAL BEST	28	.076	.055	.131	1.38	.084	.045	.129	-.008	.010
HARD42	PROP CORRECT	27	.122	.025	.147	4.88	.120	.028	.148	.002	-.003
	PROP CORRECT	28	.101	.038	.139	2.66	.084	.045	.129	.017	-.007
	PROP CORRECT	29	.076	.068	.144	1.12	.055	.069	.124	.021	-.001
	BINOMIAL ERROR	29	.076	.068	.144	1.12	.067	.059	.126	.009	.009
	BAYES	29	.076	.068	.144	1.12	.079	.055	.134	-.003	.013
	EMPIRICAL BEST	28	.101	.038	.139	2.66	.084	.045	.129	.017	-.007
HARD43	PROP CORRECT	27	.228	.038	.266	6.00	.120	.028	.148	.108	.010
	PROP CORRECT	28	.173	.055	.228	3.15	.084	.045	.129	.089	.010
	PROP CORRECT	29	.135	.072	.207	1.88	.055	.069	.124	.080	.003
	BINOMIAL ERROR	29	.135	.072	.207	1.88	.061	.074	.135	.074	-.002
	BAYES	29	.135	.072	.207	1.88	.079	.063	.142	.056	.009
	EMPIRICAL BEST	32	.051	.143	.194	.357	.009	.167	.176	.042	-.024

Table C (cont): 40 Round Hard Tests and 120 Round Hard Criterion



SUBTEST	MODEL	SCORE	OBSERVED				EXPECTED				DIFFERENCE		
			MISCLASSIFICATION		MISCLASSIFICATION		MISCLASSIFICATION		MISCLASSIFICATION		FP	FN	TOT
TABLE15	PROP CORRECT	7	.004	.110	.114	.036	.009	.035	.044	-.005	.075	.070	
	PROP CORRECT	8	.004	.186	.190	.022	.004	.105	.109	0	.081	.081	
	BINOMIAL ERROR	6	.008	.072	.080	.111	.086	.007	.093	-.078	.065	-.013	
	BAYES	7	.004	.110	.114	.036	.103	.031	.134	-.099	.079	-.020	
	EMPIRICAL BEST 0-2		.021	0	.021	UND	.021	0	.021	0	0	0	
TABLE16	PROP CORRECT	7	.004	.177	.181	.023	.009	.035	.044	-.005	.142	.137	
	PROP CORRECT	8	0	.266	.266	0	.004	.105	.109	-.004	.161	.157	
	BINOMIAL ERROR	7	.004	.177	.181	.023	.065	.060	.125	-.061	.117	.056	
	BAYES	7	.004	.177	.181	.023	.090	.054	.144	-.086	.123	.037	
	EMPIRICAL BEST 0-2		.021	0	.021	UND	.021	0	.021	0	0	0	
TABLE17	PROP CORRECT	7	0	.042	.042	0	.009	.035	.044	-.009	.007	-.002	
	PROP CORRECT	8	0	.097	.097	0	.004	.105	.109	-.004	-.008	-.012	
	BINOMIAL ERROR	6	0	.013	.013	0	.032	.006	.038	-.032	.007	-.025	
	BAYES	7	0	.042	.042	0	.070	.016	.086	-.070	.026	-.044	
	EMPIRICAL BEST	5	.008	0	.008	UND	.018	.002	.020	-.010	-.002	-.012	
TABLE18	PROP CORRECT	7	.017	.004	.021	4.25	.009	.035	.044	.008	-.031	-.023	
	PROP CORRECT	8	.015	.021	.034	.619	.004	.105	.109	.009	-.084	-.075	
	BINOMIAL ERROR	5	.021	0	.021	UND	.002	0	.002	.019	0	.019	
	BAYES	7	.017	.004	.021	4.25	.031	.001	.032	-.014	.003	-.011	
	EMPIRICAL BEST 0-3		.021	0	.021	UND	.021	0	.021	0	0	0	
	EMPIRICAL BEST	4	.021	0	.021	UND	.020	0	.020	.001	0	.001	
	EMPIRICAL BEST	5	.021	0	.021	UND	.018	.002	.020	.003	-.002	.001	
EMPIRICAL BEST	6	.017	.004	.021	4.25	.014	.010	.024	.003	-.006	-.003		
EMPIRICAL BEST	7	.017	.004	.021	4.25	.009	.035	.044	.008	-.031	-.023		

Table C (cont): 10 Round Easy Subtests and 120 Round Easy Criterion

SUBTEST	MODEL	SCORE	OBSERVED MISCLASSIFICATION			EXPECTED MISCLASSIFICATION			DIFFERENCE			
			FP	FN	TOT	FP/FN	FP	FN	TOT	FP	FN	TOT
TABLE25	PROP CORRECT	7	0	.114	.114	0	.009	.035	.044	-.009	.079	.070
	PROP CORRECT	8	0	.232	.232	0	.004	.105	.109	-.004	.127	.123
	BINOMIAL ERROR	7	0	.114	.114	0	.063	.037	.100	-.063	.077	.014
	BAYES	7	0	.114	.114	0	.102	.032	.134	-.102	.082	-.020
	EMPIRICAL BEST	0-2	.021	0	.021	UND	.021	0	.021	0	0	0
TABLE26	PROP CORRECT	7	.004	.152	.156	.026	.009	.035	.044	-.005	.117	.112
	PROP CORRECT	8	0	.262	.262	0	.004	.105	.109	-.004	.157	.153
	BINOMIAL ERROR	7	.004	.152	.156	.026	.080	.028	.108	-.076	.124	.048
	BAYES	7	.004	.152	.156	.026	.097	.036	.133	-.093	.116	.023
	EMPIRICAL BEST	1	.017	0	.017	UND	.021	0	.021	-.004	0	-.004
TABLE27	PROP CORRECT	7	.004	.055	.059	.073	.009	.035	.044	-.005	.020	.015
	PROP CORRECT	8	.004	.114	.118	.035	.004	.105	.109	0	.009	.009
	BINOMIAL ERROR	7	.004	.055	.059	.073	.040	.014	.054	-.036	.041	.005
	BAYES	7	.004	.055	.059	.073	.061	.016	.077	-.057	.039	-.018
	EMPIRICAL BEST	4	.008	.008	.016	1.00	.020	0	.020	-.012	.008	-.004
TABLE28	PROP CORRECT	7	.017	.008	.025	2.13	.009	.035	.044	.008	-.027	-.019
	PROP CORRECT	8	.017	.013	.030	1.31	.004	.105	.109	.013	-.092	-.079
	BINOMIAL ERROR	4	.021	0	.021	UND	0	0	0	.021	0	.021
	BAYES	7	.017	.008	.025	2.13	.024	.005	.029	-.007	.003	-.004
	EMPIRICAL BEST	0-3	.021	0	.021	UND	.021	0	.021	0	0	0
	EMPIRICAL BEST	4	.021	0	.021	UND	.020	0	.020	.001	0	.001
	EMPIRICAL BEST	5	.021	0	.021	UND	.018	.002	.020	.003	-.002	.001
	EMPIRICAL BEST	6	.021	0	.021	UND	.014	.010	.024	.007	-.010	-.003

Table C (cont): 10 Round Easy Subtests and 120 Round Easy Criterion

SUBTEST	MODEL	SCORE	OBSERVED			MISCLASSIFICATION			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP/FN	FP	FN	TOT	FP	FN	TOT	FP	FN
TABLE35	PROP CORRECT	7	.017	.076	.093	.224	.009	.035	.044	.008	.041	.049		
	PROP CORRECT	8	.013	.160	.173	.081	.004	.105	.109	.009	.055	.064		
	BINOMIAL ERROR	6	.017	.038	.055	.447	.044	.012	.056	-.027	.026	-.001		
	BAYES	7	.017	.076	.093	.224	.091	.021	.112	-.074	.055	-.019		
	EMPIRICAL BEST	0-2	.021	0	.021	UND	.021	0	.021	0	0	0		
TABLE36	PROP CORRECT	7	.017	.110	.127	.155	.009	.035	.044	.008	.075	.083		
	PROP CORRECT	8	.008	.190	.198	.042	.004	.105	.109	.004	.085	.089		
	BINOMIAL ERROR	7	.017	.110	.127	.155	.049	.031	.080	-.032	.079	.047		
	BAYES	7	.017	.110	.127	.155	.087	.027	.114	-.070	.083	.013		
	EMPIRICAL BEST	0-2	.021	0	.021	UND	.021	0	.021	0	0	0		
TABLE37	PROP CORRECT	7	.013	.025	.038	.520	.009	.035	.044	.004	-.010	-.006		
	PROP CORRECT	8	.013	.089	.102	.146	.004	.105	.109	.009	-.016	-.007		
	BINOMIAL ERROR	6	.013	.008	.021	1.63	.021	.002	.023	-.008	.006	-.002		
	BAYES	7	.013	.025	.038	.520	.065	.009	.074	-.052	.016	-.036		
	EMPIRICAL BEST	0-2	.021	0	.021	UND	.021	0	.021	0	0	0		
TABLE38	EMPIRICAL BEST	4	.017	.004	.021	4.25	.020	0	.020	-.003	.004	.001		
	EMPIRICAL BEST	5	.017	.004	.021	4.25	.018	.002	.020	-.001	.002	.001		
	EMPIRICAL BEST	6	.013	.008	.021	1.63	.014	.010	.024	-.001	-.002	-.003		
	PROP CORRECT	7	.021	.004	.025	5.25	.009	.035	.044	.012	-.031	-.019		
	PROP CORRECT	8	.021	.017	.038	1.24	.004	.105	.109	.017	-.088	-.071		
TABLE39	BINOMIAL ERROR	4	.021	0	.021	UND	0	0	0	.021	0	.021		
	BAYES	7	.021	.004	.025	5.25	.027	.001	.028	-.006	.003	-.003		
	EMPIRICAL BEST	0-3	.021	0	.021	UND	.021	0	.021	0	0	0		
	EMPIRICAL BEST	4	.021	0	.021	UND	.020	0	.020	.001	0	.001		
	EMPIRICAL BEST	5	.021	0	.021	UND	.018	.002	.020	.003	-.002	.001		

Table C (cont): 10 Round Easy Subtests and 120 Round Easy Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE			
			MISCLASSIFICATION			MISCLASSIFICATION						
			FP	FN	TOT	FP/FN	FP	FN	TOT	FP	FN	TOT
EASY21	PROP CORRECT	14	0	.038	.038	0	.006	.024	.030	-.006	.014	.008
	PROP CORRECT	15	0	.076	.076	0	.003	.053	.056	-.003	.023	.020
	BINOMIAL ERROR	13	.004	.021	.025	.190	.019	.008	.027	-.015	.013	-.002
	BAYES	14	0	.038	.038	0	.052	.011	.063	-.052	.027	-.025
	EMPIRICAL BEST	11	.004	.004	.008	1.00	.016	.001	.017	-.012	.003	-.009
EASY22	PROP CORRECT	14	.004	.051	.055	.078	.006	.024	.030	-.002	.027	.025
	PROP CORRECT	15	.004	.080	.084	.050	.003	.053	.056	.001	.027	.028
	BINOMIAL ERROR	12	.017	.008	.025	2.13	.021	.003	.024	-.004	.005	.001
	BAYES	14	.004	.051	.055	.078	.049	.019	.068	-.045	.032	-.013
	EMPIRICAL BEST	9	.017	0	.017	UND	.020	0	.020	-.003	0	-.003
EASY23	EMPIRICAL BEST	10	.017	0	.017	UND	.019	0	.019	-.002	0	-.002
	PROP CORRECT	14	0	.034	.034	0	.006	.024	.030	-.006	.010	.004
	PROP CORRECT	15	0	.063	.063	0	.003	.053	.056	-.003	.010	.007
	BINOMIAL ERROR	12	.008	.004	.012	2.00	.019	.002	.021	-.011	.002	-.009
	BAYES	14	0	.034	.034	0	.054	.014	.068	-.054	.020	-.034
EASY23	EMPIRICAL BEST	13	.004	.004	.008	1.00	.010	.010	.020	-.006	-.006	-.012

Table C (cont): 20 Round Easy Subtests and 120 Round Easy Criterion

SUBTEST	MODEL	SCORE	OBSERVED MISCLASSIFICATION				EXPECTED MISCLASSIFICATION				DIFFERENCE		
			FP	FN	TOT	FP/FN	FP	FN	TOT	FP	FN	TOT	
EASY24	PROP CORRECT	14	0	.055	.055	0	.006	.024	.030	-.006	.031	.025	
	PROP CORRECT	15	0	.084	.084	0	.003	.053	.056	-.003	.010	.007	
	BINOMIAL ERROR	13	0	.025	.025	0	.023	.013	.036	-.023	.012	-.011	
	BAYES	14	0	.055	.055	0	.049	.017	.066	-.049	.038	-.011	
	EMPIRICAL BEST	10	.004	.004	.008	1.00	.019	0	.019	-.015	.004	-.011	
	EMPIRICAL BEST	12	.004	.004	.008	1.00	.013	.004	.017	-.009	0	-.009	
EASY25	PROP CORRECT	14	.017	.034	.051	.500	.006	.024	.030	.011	.010	.021	
	PROP CORRECT	15	.017	.063	.080	.270	.003	.053	.056	.014	.010	.024	
	BINOMIAL ERROR	12	.021	0	.021	UND	.013	0	.013	.008	0	.008	
	BAYES	14	.017	.034	.051	.500	.044	.011	.055	-.027	.023	-.004	
	EMPIRICAL BEST	0-8	.021	0	.021	UND	.021	0	.021	0	0	0	
	EMPIRICAL BEST	9	.021	0	.021	UND	.020	0	.020	.001	0	.001	
	EMPIRICAL BEST	10	.021	0	.021	UND	.019	0	.019	.002	0	.002	
	EMPIRICAL BEST	11	.021	0	.021	UND	.016	.001	.017	.005	-.001	.004	
	EMPIRICAL BEST	12	.021	0	.021	UND	.013	.004	.017	.008	-.004	.004	
	EMPIRICAL BEST	14	.021	.008	.029	2.63	.006	.024	.030	.015	-.016	-.001	
EASY26	PROP CORRECT	15	.017	.046	.063	2.71	.003	.053	.056	.014	-.007	.007	
	BINOMIAL ERROR	11	.021	0	.021	UND	.003	0	.003	.018	0	.018	
	BAYES	14	.021	.008	.029	2.62	.053	.002	.055	-.032	.006	-.026	
	EMPIRICAL BEST	0-8	.021	0	.021	UND	.021	0	.021	0	0	0	
	EMPIRICAL BEST	9	.021	0	.021	UND	.020	0	.020	.001	0	.001	
	EMPIRICAL BEST	10	.021	0	.021	UND	.019	0	.019	.002	0	.002	
	EMPIRICAL BEST	11	.021	0	.021	UND	.016	.001	.017	.005	-.001	.004	
	EMPIRICAL BEST	12	.021	0	.021	UND	.013	.004	.017	.005	-.004	.004	
	EMPIRICAL BEST	14	.021	.008	.029	2.63	.006	.024	.030	.015	-.016	-.001	
	EMPIRICAL BEST	15	.017	.046	.063	2.71	.003	.053	.056	.014	-.007	.007	

Table C (cont): 20 Round Easy Subtests and 120 Round Easy Criterion

SUBTEST	MODEL	SCORE	OBSERVED			EXPECTED			DIFFERENCE		
			FP	FN	TOT	FP	FN	TOT	FP	FN	TOT
EASY41	PROP CORRECT	27	0	.008	.008	.006	.008	.014	-.006	0	-.006
	PROP CORRECT	28	0	.038	.038	.004	.015	.019	-.004	.023	.019
	PROP CORRECT	29	0	.059	.059	.002	.026	.028	-.002	.033	.031
	BINOMIAL ERROR	26	0	.004	.004	.014	.003	.017	-.014	.001	-.013
	BAYES	28	0	.038	.038	.026	.014	.040	-.026	.024	-.002
	EMPIRICAL BEST	26	0	.004	.004	.008	.004	.012	-.008	0	-.008
EASY42	PROP CORRECT	27	0	.021	.021	.006	.008	.014	-.006	.013	.007
	PROP CORRECT	28	0	.025	.025	.004	.015	.019	-.004	.010	.006
	PROP CORRECT	29	0	.038	.038	.002	.026	.028	-.002	.012	.010
	BINOMIAL ERROR	27	0	.021	.021	.010	.009	.019	-.010	.012	.002
	BAYES	28	0	.025	.025	.037	.007	.044	-.037	.018	-.019
	EMPIRICAL BEST	25	0	0	0	.010	.002	.012	-.010	-.002	-.012
EASY43	PROP CORRECT	27	.021	.008	.029	.006	.008	.014	.015	0	.015
	PROP CORRECT	28	.017	.017	.034	.004	.015	.019	.013	.002	.015
	PROP CORRECT	29	.017	.030	.047	.002	.026	.028	.015	.004	.019
	BINOMIAL ERROR	25	.021	0	.021	.007	0	.007	.014	0	.014
	BAYES	28	.017	.017	.034	.026	.007	.033	-.009	.010	.001
	EMPIRICAL BEST	0-18	.021	0	.021	.021	0	.021	0	0	0
	EMPIRICAL BEST	19-20	.021	0	.021	.020	0	.020	.001	0	.001
	EMPIRICAL BEST	21	.021	0	.021	.018	0	.018	.003	0	.003
	EMPIRICAL BEST	22	.021	0	.021	.017	0	.017	.004	0	.004
	EMPIRICAL BEST	23	.021	0	.021	.015	0	.015	.006	0	.006
EASY43	EMPIRICAL BEST	24	.021	0	.021	.013	.001	.014	.008	-.001	.007
	EMPIRICAL BEST	25	.021	0	.021	.010	.002	.012	.011	-.002	.009

Table C (cont): 40 Round Easy Subtests and 120 Round Easy Criterion

SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY	SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY
TABLE11	PROP CORRECT BINOMIAL ERROR BAYES	8.775 4.638 3.864	6.925 6.913 -4.236	TABLE12	PROP CORRECT BINOMIAL ERROR BAYES	7.585 3.645 3.746	-4.875 -4.886 -10.723
TABLE13	PROP CORRECT BINOMIAL ERROR BAYES	10.693 6.063 5.203	-20.075 -20.084 -19.080	TABLE14	PROP CORRECT BINOMIAL ERROR BAYES	7.332 3.551 3.242	3.325 2.314 -6.765
TABLE21	PROP CORRECT BINOMIAL ERROR BAYES	6.467 3.255 2.656	9.225 9.208 -2.971	TABLE22	PROP CORRECT BINOMIAL ERROR BAYES	7.535 3.660 3.289	1.225 1.208 -7.370
TABLE23	PROP CORRECT BINOMIAL ERROR BAYES	12.005 7.289 5.624	-25.275 -25.287 -21.940	TABLE24	PROP CORRECT BINOMIAL ERROR BAYES	9.037 4.898 3.575	6.125 6.110 -4.675
TABLE31	PROP CORRECT BINOMIAL ERROR BAYES	8.365 4.270 3.504	2.425 2.412 -6.710	TABLE32	PROP CORRECT BINOMIAL ERROR BAYES	9.355 5.196 4.050	10.825 10.815 -2.091
TABLE33	PROP CORRECT BINOMIAL ERROR BAYES	10.361 5.527 4.805	-5.675 -5.690 -11.163	TABLE34	PROP CORRECT BINOMIAL ERROR BAYES	10.965 6.722 4.451	16.825 16.815 1.208

Table D : Average Per Test Sum of Squared and Absolute Discrepancies Between Estimated True Scores and Criterion True Scores: 10 Round Hard Subtests and 120 Round Hard Criterion

SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY	SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY
HARD21	PROP CORRECT	3.526	-.475	HARD22	PROP CORRECT	4.197	-7.375
	BINOMIAL ERROR	2.254	-.423		BINOMIAL ERROR	2.533	2.334
	BAYES	2.410	-5.527		BAYES	2.778	-10.422
HARD23	PROP CORRECT	3.603	-.225	HARD24	PROP CORRECT	3.648	-4.125
	BINOMIAL ERROR	2.511	7.347		BINOMIAL ERROR	2.334	4.663
	BAYES	2.181	-5.349		BAYES	2.308	-8.116
HARD25	PROP CORRECT	4.685	7.875	HARD26	PROP CORRECT	4.426	4.325
	BINOMIAL ERROR	3.133	7.863		BINOMIAL ERROR	2.909	4.315
	BAYES	2.771	.398		BAYES	2.844	-2.121
HARD41	PROP CORRECT	2.184	-3.925	HARD42	PROP CORRECT	1.952	-2.175
	BINOMIAL ERROR	1.655	-3.936		BINOMIAL ERROR	1.400	-2.187
	BAYES	1.769	-6.294		BAYES	1.451	-4.841
HARD43	PROP CORRECT	3.150	6.100				
	BINOMIAL ERROR	2.494	6.088				
	BAYES	2.341	2.028				

Table D (cont): 20 Round and 40 Round Subtests and 120 Round Hard Criterion



SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY	SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY
TABLE15	PROP CORRECT BINOMIAL ERROR BAYES	5.144 2.256 4.046	-13.492 -13.496 -19.578	TABLE16	PROP CORRECT BINOMIAL ERROR BAYES	6.395 3.131 4.871	-16.692 -16.701 -21.984
TABLE17	PROP CORRECT BINOMIAL ERROR BAYES	2.617 1.091 1.697	2.408 2.396 -7.625	TABLE18	PROP CORRECT BINOMIAL ERROR BAYES	3.184 2.662 1.569	17.808 17.794 3.952
TABLE25	PROP CORRECT BINOMIAL ERROR BAYES	5.034 2.255 3.766	-11.692 -11.706 -18.225	TABLE26	PROP CORRECT BINOMIAL ERROR BAYES	7.577 4.273 5.332	-16.992 -17.003 -22.209
TABLE27	PROP CORRECT BINOMIAL ERROR BAYES	3.985 2.177 2.422	2.708 2.692 -7.399	TABLE28	PROP CORRECT BINOMIAL ERROR BAYES	3.194 2.800 1.545	18.808 18.786 4.704
TABLE35	PROP CORRECT BINOMIAL ERROR BAYES	3.542 1.404 2.646	-3.192 -3.208 -11.835	TABLE36	PROP CORRECT BINOMIAL ERROR BAYES	5.099 2.405 3.667	-6.892 -6.905 -14.616
TABLE37	PROP CORRECT BINOMIAL ERROR BAYES	3.149 1.764 1.885	7.908 7.894 -3.490	TABLE38	PROP CORRECT BINOMIAL ERROR BAYES	3.570 3.030 1.795	19.308 19.283 5.080

Table D (cont): 10 Round Easy Subtests and 120 Round Easy Criterion

SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY	SUBTEST	MODEL	SQUARED DISCREPANCY	ABSOLUTE DISCREPANCY
EASY21	PROP CORRECT	1.647	-2.792	EASY22	PROP CORRECT	1.418	-2.192
	BINOMIAL ERROR	.851	-2.806		BINOMIAL ERROR	.765	-2.205
	BAYES	1.445	-7.782		BAYES	1.287	-7.267
EASY23	PROP CORRECT	1.084	-.842	EASY24	PROP CORRECT	1.401	-2.742
	BINOMIAL ERROR	.536	-.850		BINOMIAL ERROR	.670	-2.754
	BAYES	.931	-6.109		BAYES	1.209	-7.739
EASY25	PROP CORRECT	1.514	3.858	EASY26	PROP CORRECT	1.299	4.708
	BINOMIAL ERROR	.963	3.849		BINOMIAL ERROR	.924	4.698
	BAYES	1.187	-2.075		BAYES	1.010	-1.345
EASY41	PROP CORRECT	.702	-2.492	EASY42	PROP CORRECT	.675	-1.792
	BINOMIAL ERROR	.477	-2.501		BINOMIAL ERROR	.404	-1.799
	BAYES	.693	-5.200		BAYES	.622	-4.554
EASY43	PROP CORRECT	.939	4.283				
	BINOMIAL ERROR	.769	4.270				
	BAYES	.794	1.058				

Table D (cont): 20 Round and 40 Round Subtests and 120 Round Easy Criterion

## REFERENCES

- Baker, E.L. Beyond objectives: Domain-Referenced tests for evaluation and instructional improvement. Educational Technology, 1974, 14, 6, 10-17.
- Block, J.H. Student learning and the setting of mastery performance standards. Educational Horizons, 1972, 50, 183-190.
- Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnam, N. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley, 1972.
- Davis, F.B. 1971 AERA Conference Summaries: II. Criterion-referenced measurement. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1972.
- Dawes, R.M. and Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.
- Dayton, C.M. and Macready, G.B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.
- Dodd, D.H. and Schultz, R.F. Computational procedures for estimating magnitude of effect for some analysis of variance designs. Psychological Bulletin, 1973, 79, 392-395.
- Donlon, T.F. Some needs for clearer terminology in criterion referenced testing. Paper presented at the annual meeting of the National Council for Measurement in Education, Chicago, 1974.
- Ebel, R.L. Content standard scores. Educational and Psychological Measurement. 1962, 22, 15-25.
- Harick, J.A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326. (a)
- Harick, J.A. The experimental validation of an evaluation model for mastery testing: Final report, Project No. O-A-073. Washington, D.C.: Office of Education, U.S. Department of Health, Education, and Welfare, 1971. (b)
- Harick, J.A. and Adams, E.N. An evaluation model for individualized instruction. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970.

- Epstein, K.I. A generalization of the Emrick model for the case of unequal proportions of masters and nonmasters. In F.H. Steinheiser, Jr., K.I. Epstein, A. Mirabella, and G.B. Macready, Criterion-referenced testing: A critical analysis of selected models: ARI Technical Paper 306. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, August 1978.
- Epstein, K.I. An empirical investigation of four criterion-referenced testing models. Paper presented at the 17th Annual Conference of the Military Testing Association, Indianapolis, 1975.
- Epstein, K.I., Knerr, C.S. Criterion-Referenced test interpretations of "classical" measurement theory. A paper presented at the American Educational Research Association, annual meeting, San Francisco, California, April 1976 (ERIC Document No. ED126154).
- Fitts, P.M. and Posner, M.I. Human Performance. Belmont, California: Brooks/Cole, 1967.
- Gagné, R.M. and Briggs, L.J. Principles of Instructional Design. New York: Holt, Rinehart and Winston, Inc., 1974.
- Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.
- Glaser, R. and Klaus, D.J. Proficiency measurement: Assessing human performance. In Gagné, R.M. (Ed.) Psychological principles in system development. New York: Holt, Rinehart and Winston, 1963.
- Glass, G.V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.
- Graham, D.L. An empirical investigation of the application of criterion-referenced measurement to survey achievement testing. Unpublished doctoral dissertation, Tallahassee, Florida: Florida State University, 1974.
- Graham, D.L. and Bergquist, C.C. An examination of criterion-referenced test characteristics in relation to assumptions about the nature of achievement variables. Paper presented at the annual meeting of the American Educational Research Association, Washington, 1975.
- Hambleton, R.K. and Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.
- Hambleton, R.K., Swaminathan, H., Algina, J., and Coulson, D.B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

- Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., and Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.
- Hambleton, R.K. and Traub, R.E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.
- Hively, W., Patterson, H.L., and Page, S.H. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.
- Kifer, E. and Bramble, W. The calibration of a criterion-referenced test. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974.
- Kriewall, T.E. Application of information theory and acceptance sampling principles to the management of mathematics instruction. Technical Report No. 103. Madison, Wisconsin: Wisconsin Research and Development Center, 1969.
- Kriewall, T.E. Aspects and applications of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.
- Lord, F.M. and Novick, M.R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Company, 1968.
- Macready, G.B., and Dayton, C. The use of probabilistic models in the assessment of mastery. Unpublished manuscript, College Park, Maryland: University of Maryland, 1975.
- Merrill, M.D. Psychomotor and Memorization Behavior. In Merrill, M.D. (Ed.) Instructional design: Readings. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1971.
- Millman, J. Determining test length. Los Angeles, California: Instructional Objectives Exchange, 1972.
- Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-215.
- Millman, J. Criterion-referenced measurement. In Popham, W.J. (Ed.) Evaluation in education: Current applications, Berkeley, California: McCutchen Publishing Corporation, 1974.
- Millman, J. Hang the hang-ups about test making. Paper presented at the First Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C., October 27, 1978.

- National Council on Measurement in Education, Inc. Journal of Educational Measurement, 1977, 14, 73-196.
- Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., and Bent, D.M. Statistical Package for the Social Sciences (2nd Edition). New York: McGraw-Hill Book Company, 1975.
- Novick, M.R. High school attainment: An example of a computer-assisted Bayesian approach to data analysis. International Statistical Review, 1973, 41, 264-271.
- Novick, M.R. and Jackson, P.H. Statistical Methods for Educational and Psychological Research. New York: McGraw-Hill Book Company, 1974.
- Novick, M.R. and Lewis, C. Prescribing test length for criterion-referenced measurement. In C.W. Harris, M.C. Alkin, and W.J. Popham (Eds.), Problems in criterion-referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.
- Novick, M.R., Lewis, D., and Jackson, P.H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-46.
- Popham, W.J. A lasso for runaway test items. Paper presented at the First Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C., October 27, 1978.
- Roudabush, G.E. Models for a beginning theory of criterion-referenced tests. Paper presented at the annual meeting of the National Council for Measurement in Education, Chicago, 1974.
- Steinheiser, F.H. Jr. and Epstein, K.I. Analysis of variance: Selection of a model and summary statistics. Paper presented at the 23rd Conference on the Design of Experiments in Army Research, Development and Testing, Monterey, California, October 1977.
- Steinheiser, F.H. Jr. and Epstein, K.I. An experimental investigation of the Military Police Firearms Qualification Course: ARI Technical Paper 322. Alexandria, Virginia: US Army Research Institute for the Behavioral and Social Sciences, September 1978.
- Steinheiser, F.H. Jr., Epstein, K.I., Mirabella, A., and Macready, G.B. Criterion-referenced testing: A critical analysis of selected models: ARI Technical Paper 306. Alexandria, Virginia: US Army Research Institute for the Behavioral and Social Sciences, August 1978.
- US Army Military Police School. Army Training Circular 19-4. Fort McClelland, Alabama, 1975.

- Wilcox, R. and Harris, C.W. On Earick's "An evaluation model for mastery testing". Journal of Educational Measurement, 1977, 14, 215-218.
- Winer, B.J. Statistical Principles in Experimental Design (2nd Edition). New York: McGraw-Hill Book Company, 1971.
- Wright, B.D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1968, 85-101.
- Wright, B.D., and Mead, R.J. CALFIT: Sample-free item calibration with a Rasch measurement model: Research Memorandum No. 18. Chicago, Illinois: Statistical Laboratory, Department of Education, University of Chicago, 1975.
- Wright, B.D., and Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.
- Whitely, S.E. and Dawis, R.V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 163-178.